UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA TÉCNICA DE TELECOMUNICACIÓN
SONIDO E IMAGEN

PROYECTO FINAL DE CARRERA

ESTUDIO COMPARATIVO DE PARÁMETROS ESPECTRALES PARA CLASIFICACIÓN DE AUDIO

AUTOR: ENRIQUE PRIETO LABRADOR
TUTORA: ASCENSIÓN GALLARDO ANTOLÍN

30 DE MAYO DE 2008

TÍTULO: *ESTUDIO COMPARATIVO DE PARÁMETROS ESPECTRALES PARA CLASIFICACIÓN DE AUDIO*

AUTOR: *ENRIQUE PRIETO LABRADOR*

TUTORA: *ASCENSIÓN GALLARDO ANTOLÍN*

La defensa del presente Proyecto Fin de Carrera se realizó el día 30 de Mayo de 2008; siendo calificada por el siguiente tribunal:

PRESIDENTE: *FERNANDO DÍAZ DE MARÍA*

SECRETARIO: *RUBÉN SOLERA UREÑA*

VOCAL: *CARLOS GARCÍA RUBIO*

Habiendo obtenido la siguiente calificación:

CALIFICACIÓN:

Presidente **Secretario** **Vocal**

Agradecimientos

A mis amigos que siempre han estado cuando les he necesitado, y en estos últimos tiempos más que nunca. A Teresa, por lo mismo y más aún.

A mi tutora Ascen, sin la que no podría haber terminado nunca este proyecto.

Especialmente a mi familia, a mis padres. Espero que esto les sirva de alegría en momentos difíciles, y que vean un pequeño fruto de lo que me han dado, que es todo.

Y sobre todo a la persona más especial que compartió mi vida durante más de 22 años, y que siempre estará conmigo. Va por ti Edu.

Nunca hay un adios total
entre dos ñeris,
siempre es nos volveremos a ver
en algun lugar del tiempo

No hay olvido cuando existe
la amistad y el respeto
El recuerdo de momentos entrañables
de alegrías y secretos

Resumen

Este Proyecto de Fin de Carrera se enmarca dentro de un conjunto de líneas de investigación sobre el procesado de registros de audio que lleva a cabo el Departamento de Teoría de la Señal y Comunicaciones de la Universidad Carlos III de Madrid.

El trabajo se ha centrado en una de esas líneas, teniendo como objetivo evaluar el rendimiento de un sistema de clasificación de audio con diferentes parámetros utilizados habitualmente para la caracterización de la señal de audio, como son los MFCC y los ASE y ASP definidos en el estándar MPEG-7.

Se han realizado una serie de experimentos con diferentes configuraciones del módulo de parametrización con el objeto de determinar las características acústicas que permiten una mejor clasificación del audio en las tres clases propuestas para este trabajo: habla, música instrumental y música con voz.

Índice general

Lista de figuras

Lista de tablas

1 Introducción

La intención de este proyecto es hallar un conjunto de características propias de la señal de audio que permita diferenciar de la mejor forma posible los diferentes tipos de archivos de audio. Para este proyecto se quieren diferenciar tres tipos de audio: habla, música instrumental, y música con voz.

1.1 Definición del problema

Cada vez, la necesidad de tener colecciones de archivos estructuradas es mayor, ante la gran cantidad de información que manejamos habitualmente, debido a la mayor fluidez de las comunicaciones y transmisión de datos, y a las grandes capacidades de almacenamiento de los equipos modernos. Esta necesidad se puede hacer mucho más manifiesta en entornos como servidores de información, y más concretamente en nuestro caso, de información de audio.

Una estructuración básica de los archivos de audio sería la diferenciación entre archivos que sólo contienen habla, archivos que sólo contienen música instrumental, y archivos que contienen ambas cosas. Esta división nos puede permitir, además, diferenciar contenidos radiofónicos, por ejemplo.

Pero, ¿cómo hacer esa diferenciación?, ¿qué características de estos archivos de audio nos permiten diferenciar un tipo de información respecto de otro? En este proyecto trataremos de descubrir qué características del audio permiten una discriminación más eficaz entre estos tipos de audio.

1.2 Objetivos

El objetivo de este proyecto es determinar experimentalmente qué parámetros de la señal de audio nos permitirán una mejor discriminación entre habla, música con voz, y música instrumental.

Hasta ahora se han utilizado frecuentemente para reconocimiento de habla, o hablante los parámetros mel-cepstrum (*Mel-Frequency Cepstrum Coefficients*, MFCC) con éxito, por lo que en primer lugar, determinaremos el funcionamiento de este tipo de características para nuestro propósito.

Recientemente con la aparición del estándar MPEG-7 se han abierto otras posibilidades a la hora de alcanzar nuestro objetivo, ya que define una serie de características que podrían ser de utilidad para obtener una mejor parametrización y, por tanto, una mejor clasificación entre las tres clases de la tarea propuesta.

Analizaremos, pues, los resultados tanto con los parámetros MFCC, como con los que se nos ofrecen en el estándar MPEG-7, utilizando diferentes configuraciones de estos últimos, y compararemos ambos, determinando qué parámetros resultan más efectivos para nuestras intenciones.

1.3 Organización del proyecto

El proyecto está estructurado en seis capítulos. El primer capítulo corresponde con esta introducción. El capítulo 2 trata sobre el módulo de parametrización, tanto el basado en parámetros MPEG-7, como el basado en parámetros MFCC. Se hará una breve descripción de todas esas características, y de cómo se extraen de los archivos de audio.

El capítulo 3 se dedica al sistema de clasificación. Se mostrarán las técnicas de clasificación más comunes, y que podrían aplicarse a este proyecto, y se explicará el sistema que se utiliza, y que está basado en los modelos de mezclas de gaussianas (*Gaussian Mixture Models*, GMM).

Una vez definida la base teórica del proyecto, en el capítulo 4 se continúa con la descripción del marco en el que se desarrollan los experimentos; básicamente se define la estructura de la base de datos sobre la que se realizarán las pruebas. También se describe en qué consisten dichos experimentos.

El capítulo 5 contiene la parte experimental del proyecto, donde se muestran los resultados de las pruebas realizadas. Por un lado se muestran los obtenidos con parametrización MFCC, y por otro lado los que resultan de la utilización de características MPEG-7, en las diferentes configuraciones consideradas. Con los resultados de todos los experimentos, haremos un análisis de los mismos, y determinamos cuales serán más útiles para nuestro fin. Para determinar cuales son mejores o peores, buscaremos los que mejores resultados de clasificación nos proporcionen y que además, no comporten una complejidad computacional demasiado elevada.

Finalmente, en el capítulo 6, a partir de los resultados obtenidos y teniendo ya las características que juzgaremos mejores para el objetivo, obtendremos unas conclusiones finales, sobre las que se podrán basar posibles líneas de investigación en un futuro, y que se determinarán al final de este proyecto.

2 Base teórica de la caracterización de audio

Este capítulo trata de mostrar algunas de las posibilidades que existen actualmente a la hora de parametrizar la señal de audio para su correcta caracterización. De esas posibilidades que se muestran a continuación, se determinará posteriormente, mediante la experimentación, cuáles dan unas mejores prestaciones para la finalidad de este proyecto, que es la correcta diferenciación entre los tres tipos de audio que se han propuesto (habla, música instrumental, y música con voz).

2.1 Introducción a la parametrización y tipos

La parametrización adecuada de la información contenida en los archivos de audio es un aspecto fundamental para la correcta clasificación de éstos en las diferentes clases acústicas consideradas. Se trata de extraer las características que mejor definen a estos archivos, y especialmente las que les hacen más diferentes a los de una clase respecto de los de otra. La búsqueda de estos parámetros que permiten una mejor separación entre clases es el objetivo de este proyecto.

Nos centraremos únicamente en los parámetros MFCC y en los definidos por el estándar MPEG-7 para audio. Los primeros han sido utilizados con éxito previamente en aplicaciones relacionadas con el audio, especialmente con el habla, y analizaremos su funcionamiento en la tarea propuesta. Los segundos ofrecen una parametrización diferente y quizá den mejor resultado que los ya mencionados MFCC.

A continuación se describen ambos tipos de parametrización.

2.2 MPEG-7 Audio

2.2.1 Introducción

El entorno de audio MPEG-7 viene a ser un conjunto de herramientas de bajo nivel diseñadas para proporcionar una base que permita construir aplicaciones de audio de mayor nivel [1, 11]. Este entorno proporciona tanto las estructuras necesarias para representar características de audio, como un conjunto básico de características.

El conjunto de características o descriptores de MPEG-7 está compuesto de 17 diferentes [2], tanto temporales, como espectrales. Se dividen en 6 grupos: *Basic*, *Basic Spectral*, *Signal Parameters*, *Timbral Temporal*, *Timbral Spectral*, y *Spectral Basis*. Además MPEG-7 también proporciona un útil Descriptor de silencios.

2.2.2 Descriptores

A continuación se enumera cada una de las características que ofrece MPEG-7, grupo a grupo:

- *Basic*: los descriptores de este grupo están muestreados a intervalos regulares de tiempo, con valores escalares. El descriptor *AudioWaveform* caracteriza la envolvente de la forma de audio (mínimo y máximo) y se utiliza frecuentemente como una representación general de la señal. El descriptor *AudioPower*, en cambio, describe la potencia instantánea temporalmente suavizada, útil como resumen compacto de la señal y se utiliza en conjunción con los descriptores del espectro de potencia.

- *Basic Spectral*: los descriptores de este grupo comparten una base común: todos derivan de un análisis sencillo en tiempo y frecuencia de la señal de audio. Todos los descriptores de este grupo se calculan a partir del descriptor *AudioSpectrumEnvelope*, que es un espectro de frecuencia

logarítmico, espaciado por un divisor potencia de dos, o múltiplo de una octava. Es un vector que describe el espectro de potencia a corto plazo. Puede usarse para representar un espectrograma, una aproximación del espectro o como descriptor de propósito general para búsqueda y comparación.

El descriptor *AudioSpectrumCentroid* describe el centro de gravedad del espectro de potencia en frecuencia logarítmica. Es una descripción económica de la forma del espectro de potencia, que indica si el contenido espectral de la señal está dominado por las altas o por las bajas frecuencias.

El descriptor *AudioSpectrumSpread* complementa al anterior, describiendo el segundo momento del espectro de potencia en frecuencia logarítmica, indicando así, si el espectro de potencia está centrado alrededor del centroide espectral, o se extiende a lo largo de todo el espectro. Esto puede ayudar a distinguir entre sonidos de tipo tonal o ruidoso.

El descriptor *AudioSpectrumFlatness* describe cómo de plano es el espectro de la señal de audio para cada una de las bandas de frecuencia para las que se mide. Cuando este vector presenta una alta desviación respecto de un espectro plano para cierta banda de frecuencia, puede ser señal de la presencia de componentes tonales.

- *SignalParameters*: los dos descriptores de este grupo se aplican principalmente a señales periódicas o cuasiperiódicas. El descriptor *AudioFundamentalFrequency* describe la frecuencia fundamental de la señal de audio. El problema de este descriptor es que no siempre es posible extraer de forma fiable la frecuencia fundamental en algunas secciones de la señal, como, por ejemplo, en las que sean ruidosas. El descriptor *AudioHarmocity* representa la armonía de la señal, y permite, así diferenciar entre señales con espectro armónico (p. ej. tonos musicales o habla sonora), con espectro inarmónico (p.ej. sonidos

metálicos), o con espectro no armónico (p. ej. ruido, habla sorda, o mezclas densas de instrumentos).

- *Timbral Temporal*: estos descriptores se basan en características temporales de segmentos de sonido, y son especialmente útiles para la descripción de timbre musical (cualidad característica de un tono, independiente del volumen y la frecuencia fundamental). Se usa un valor escalar para representar la evolución de un sonido o segmento de audio en el tiempo.

  El descriptor *LogAttackTime* caracteriza el "ataque" de un sonido, es decir, el tiempo que tarda la señal en ir desde el silencio hasta su máxima amplitud. Permite diferenciar entre un sonido brusco y uno suave.

  El descriptor *TemporalCentroid* también caracteriza la envolvente de la señal, e indica en qué segmento se concentra la energía de una señal. Permite, por ejemplo, distinguir entre una nota de piano decadente y una nota de órgano sostenida, cuando las longitudes y los ataques de ambas notas son idénticos.

- *TimbralSpectral*: estos descriptores tienen como función la descripción de la estructura armónica del espectro. Para ello, muestran características espectrales del sonido en una escala lineal de frecuencia. Son especialmente útiles para la percepción del timbre musical.

  El descriptor *SpectralCentroid* es el centro de gravedad del espectro de potencia en escala lineal. Se asocia a la "claridad" o pureza de los sonidos.

  El descriptor *AudioSpectrumCentroid* es similar al anterior descriptor, pero especializado en distinguir timbres de instrumentos musicales. Tiene una alta correlación con la característica perceptual de la agudeza de un sonido.

Los siguientes descriptores operan con las componentes armónicas de las señales, regularmente espaciadas. Por ello, se calculan en un espacio lineal de frecuencia. El descriptor *HarmonicSpectralCentroid* es la media de los picos armónicos del espectro ponderados por sus respectivas amplitudes. Sólo se aplica a las partes armónicas (no ruidosas) de la señal de audio.

El descriptor *HarmonicSpectralDeviation* indica la desviación de la amplitud de las componentes armónicas respecto a la envolvente espectral global.

El descriptor *HarmonicSpectralSpread* describe la desviación estándar de los picos armónicos del espectro ponderados por su amplitud respectiva, y normalizada por el valor del *HarmonicSpectralCentroid* instantáneo.

El descriptor *HarmonicSpectralVariation* es la correlación normalizada entre la amplitud de los picos armónicos en dos segmentos consecutivos de la señal.

- *SpectralBasis*: estos descriptores representan proyecciones de dimensión baja de un espacio espectral de dimensión alta, y son útiles para tareas de comparación y clasificación. Se usan principalmente con las Herramientas de Descripción de Indexado para Clasificación de Sonido, pero se pueden utilizar también con otros tipos de aplicaciones.

  El descriptor *AudioSpectrumBasis* es una conjunto de (potencialmente variables en el tiempo y/o estadísticamente independientes) funciones base que se derivan de la descomposición en valores singulares del espectro de potencia normalizado.

  El descriptor *AudioSpectrumProjection* se usa junto al descriptor anterior, y representa características de baja dimensión del espectro, tras

su proyección sobre una base formada por un número reducido de funciones.

Estos descriptores se pueden utilizar para ver y representar de forma compacta los subespacios independientes de un espectrograma. A menudo, estos subespacios independientes (o grupos de ellos) están correlacionados fuertemente con diferentes fuentes de sonido. Así se logra más información de la estructura del audio y es posible destacar algunos elementos del espectrograma usando una dimensión menor. Por ejemplo, en la siguiente figura, una canción pop se representa con un descriptor *AudioSpectrumEnvelope*, y se visualiza en el siguiente espectrograma:

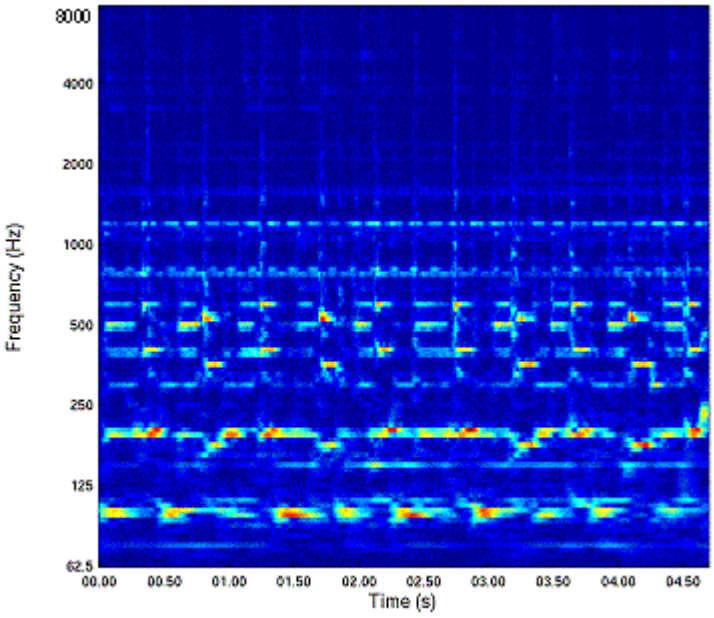

Figura 2.1: *Descripción de una canción pop con AudioSpectrumEnvelope*

Los datos que necesitamos almacenar son NM valores, siendo N el número de bandas en que dividimos el espectro, y M el número de instantes temporales en los que se analiza la señal de audio.

Es posible reducir la cantidad de datos que necesitamos almacenar para la misma canción y, además, conseguir destacar más los instrumentos individualmente con la siguiente representación en la que se utiliza una combinación de descriptores de tipo *AudioSpectrumBasis* y *AudioSpectrumProjection*:

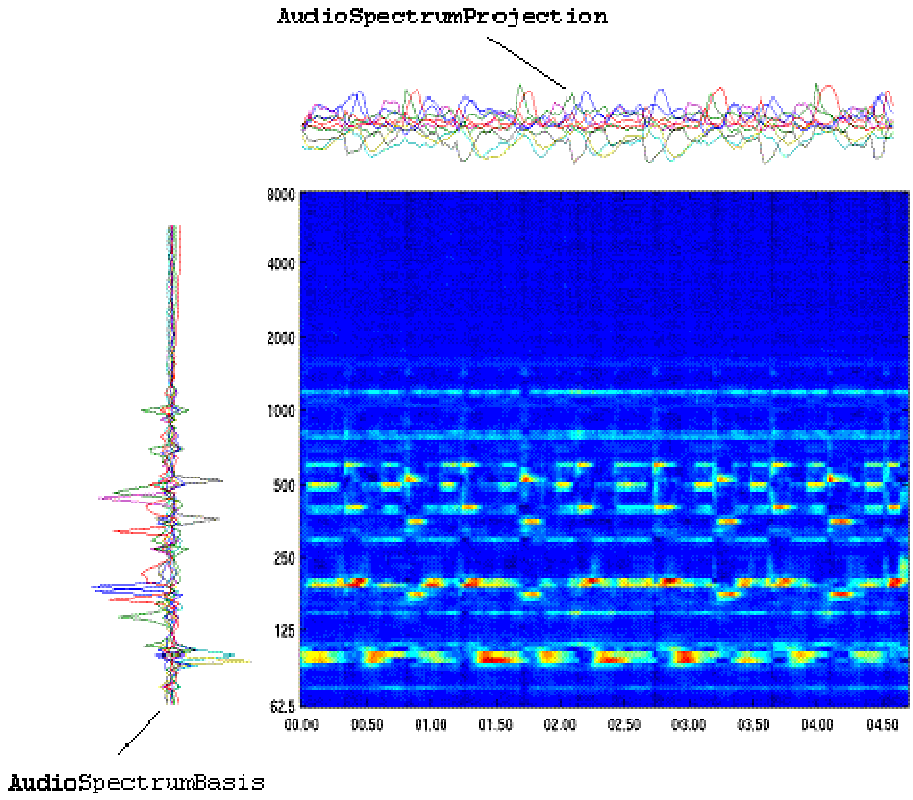

Figura 2.2: *Reconstrucción con 10 componentes base de la canción pop anterior. Se utilizan los Descriptores AudioSpectrumBasis y AudioSpectrumProjection*

Se ha utilizado una reconstrucción compuesta por 10 bases, que muestra la mayor parte del detalle del espectrograma original incluyendo las notas de guitarra, bajo, charles y órgano. Los vectores de la izquierda son un descriptor *AudioSpectrumBasis*, y los superiores corresponden al descriptor *AudioSpectrumProjection*. Con este procedimiento se necesitan almacenar 10(M+N) valores.

Existe además un descriptor para los silencios (o sonidos no significativos) denominado *SilenceSegment*. Es simple pero muy efectivo, y puede ser utilizado para ayudar a una posterior segmentación del flujo de audio, o como indicio para no procesar un segmento.

2.2.3 Parámetros de envolvente espectral (*AudioSpectrumEnvelope*, ASE)

Ahora nos centraremos únicamente en los parámetros que se utilizarán en este proyecto, y se explicarán más a fondo. Los primeros a tratar serán los parámetros de envolvente espectral (*AudioSpectrumEnvelope*, ASE) [5].

Estos parámetros, como ya hemos dicho, proporcionan información sobre la envolvente de la señal (máximos y mínimos) en cada una de las bandas en que se divide el espectro en forma logarítmica.

Se calculan de la siguiente manera:

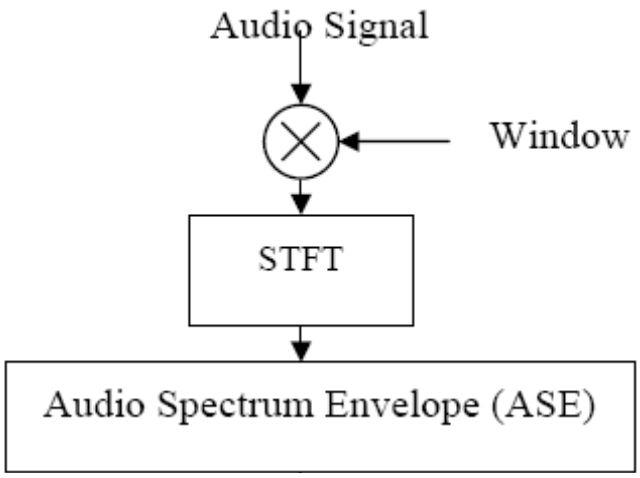

Figura 2.3: *Esquema de extracción de los parámetros ASE*

El primer paso es dividir la señal de audio $s(n)$ en tramas solapadas mediante la aplicación de un enventanado de Hamming, y analizarla usando la transformada de Fourier a corto plazo (STFT)

$$S(k,l)=\sum_{n=0}^{N-1} s(n+lM)\,w(n)\,e^{-j\frac{2nk}{N}}\,, \qquad (2.1)$$

donde k es el índice de la trama en frecuencia, l es el índice de la trama en tiempo, w es una ventana de análisis de Hamming de tamaño N, y M es el tamaño del periodo de trama. A continuación, se calcula el espectro de potencia. Por el teorema de Parseval es necesario añadir el factor de $1/N$ para equiparar la suma de las magnitudes de los coeficientes STFT con la suma de la señal enventanada al cuadrado:

$$P(k,l) = \frac{|S(k,l)|^2}{nf \cdot N} \qquad (2.2)$$

donde el factor de normalización de ventana nf se define como

$$nf = \sum_{n=0}^{N-1} w^2(n) \qquad (2.3)$$

Para reducir la dimensionalidad de las características espectrales, los coeficientes espectrales $P(k,l)$ son agrupados en sub-bandas logarítmicas. Los canales de frecuencia están espaciados de forma logarítmica en bandas de 1/4 de octava no solapadas, que se extienden desde los 62.5 Hz, que es el extremo inferior por defecto, y los 8 kHz, que es el extremo superior por defecto (habitualmente corresponde con la mitad de la frecuencia de muestreo). La salida es la suma ponderada en cada sub-banda logarítmica del espectro de potencia. En resumen, el estándar MPEG-7 define como parámetros ASE (*Audio Spectrum Envelope*) al conjunto de los siguientes elementos: un coeficiente que representa la potencia por debajo de 62.5 Hz, una serie de coeficientes que representa la potencia en bandas espaciadas logarítmicamente entre los 62.5 Hz y los 8 kHz, y un coeficiente más que representa la potencia por encima de esta última frecuencia.

Los valores de los límites de frecuencia antes mencionados son configurables. De hecho, para nuestro caso, el límite superior es 11025 Hz, ya que tomamos los ficheros de audio de la base de datos considerada están grabados a una frecuencia de muestreo de 22050 Hz.

2.2.4 Parámetros de proyección de la envolvente espectral (*Audio Spectrum Projection*, ASP)

Como ya hemos mencionado antes, los coeficientes de proyección de la envolvente espectral (*Audio Spectrum Projection*, ASP) se obtienen a partir de los ASE [5, 6], tras ser convertidos a la escala de decibelios. Cada vector espectral (en decibelios) se normaliza con el valor cuadrático medio (*Root Mean Square*, RMS) del espectro,

proporcionando así una versión normalizada en potencia logarítmica de los ASE, llamada NASE y representada por una matriz $L \times F$. Se define como:

$$X(l,f) = \frac{10\log_{10}\left(ASE(l,f)\right)}{\sqrt{\sum_{f=1}^{F}\left\{10\log_{10}\left(ASE(l,f)\right)\right\}^2}}, \qquad (2.4)$$

donde $l\left(1 \le l \le L\right)$ es el índice de trama en el tiempo, $f\left(1 \le f \le F\right)$ es el número de canales de frecuencia en escala logarítmica considerados, L es el número total de tramas y F es el número de coeficientes espectrales ASE.

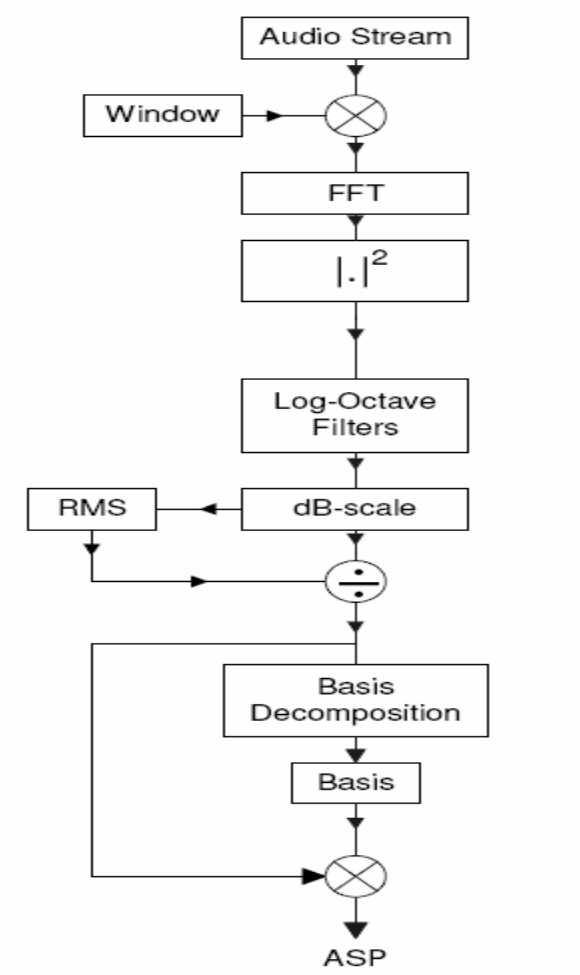

Figura 2.4: *Esquema de extracción de los parámetros ASP*

Finalmente, para calcular las características de proyección de MPEG-7, se aplica un método de descomposición en componentes principales como PCA (*Principal Component Analysis*), ICA (*Independent Component Analysis*) o NMF (*Non-negative*

Matriz Factorization). El objetivo de este procedimiento es el de reducir la dimensión del espacio de características reteniendo la mayor cantidad de información importante posible.

2.3 Parámetros mel-cepstrales (Mel-Frequency Cepstrum Coefficients, MFCC)

2.3.1 Introducción

Una gran parte de los módulos de parametrización de la señal de voz (y por extensión, de audio) se basan en el análisis espectral de ventanas temporales de la señal en las que se supone que es estacionaria o cuasi-estacionaria. Existen principalmente dos familias de parámetros, excluyendo lógicamente los que define MPEG-7, que se pueden extraer: los que obtenemos del análisis LPC (*Linear Predictive Coding*), y los MFCC (*Mel Frequency Cepstrum Coefficients*).

Para este proyecto se utilizan los parámetros MFCC [3], ya que no implican una gran carga computacional y además, producen buenos resultados en tareas de reconocimiento de habla y locutor.

El análisis cepstral es una técnica homomórfica (que cumple el principio de superposición que define a los sistemas lineales y, además, las propiedades conmutativa y distributiva, respecto de la operación de combinación de señales a la entrada) que permite separar la acción del tracto vocal, o de los instrumentos (filtro lineal variable en el tiempo) de la señal de excitación. La justificación que explica que se pueda hacer este tipo de análisis de la señal de audio se resume en la siguiente figura, aplicado al caso concreto de la voz:

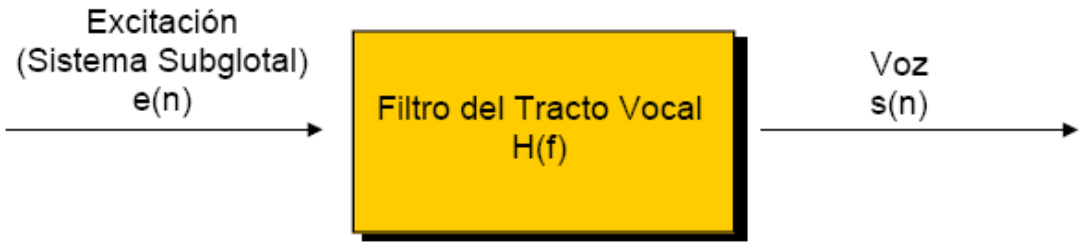

Figura 2.5: *Representación del Tracto Vocal como un filtro lineal variable en el tiempo excitado por el Sistema Subglotal*

Existe un estándar para la extracción de estos parámetros MFCC, que es el ETSI ES 201 108, y que podremos utilizar para nuestro propósito. Los coeficientes de características obtenidos se componen de 12 coeficientes cepstrales y sus primeras derivadas, y de una medida de la log-energía y su primera derivada. En total, para cada fragmento de la señal de audio, se obtiene un vector de 26 coeficientes MFCC. Aunque no se contempla en el estándar se puede realizar una operación de normalización de los parámetros MFCC.

En el siguiente diagrama se muestra el esquema de las operaciones que se realizan para el análisis cepstral de una señal de audio:

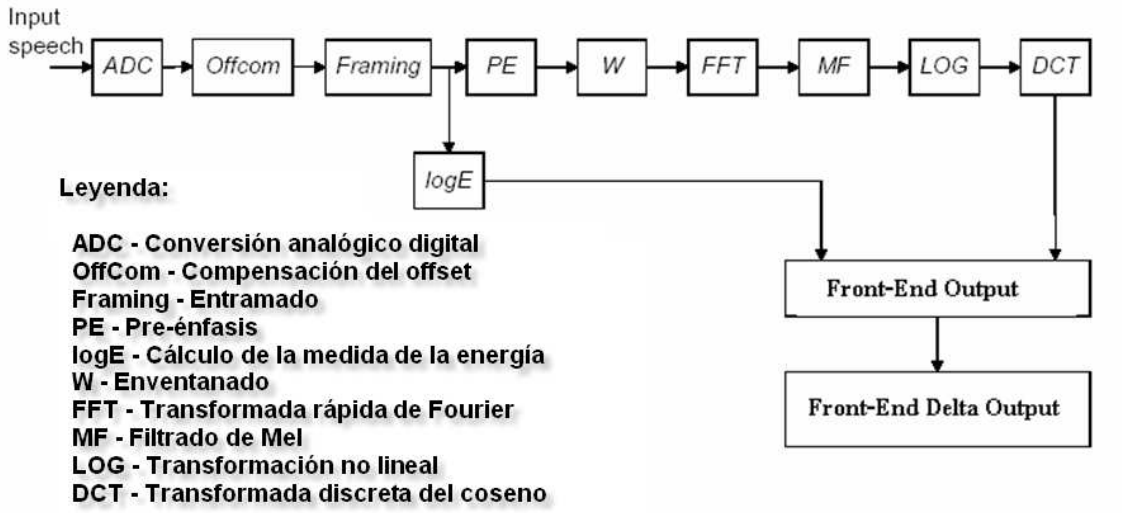

Figura 2.6: *Esquema de extracción de los parámetros MFCC*

2.3.2 Conversión analógico digital

La señal de audio es muestreada con una determinada frecuencia de muestreo, según la calidad que deseemos. En nuestro caso muestreamos con una frecuencia de 22050 Hz, y cuantificaremos con una longitud de palabra de 16 bits.

2.3.3 Compensación del offset

Se aplica un filtrado para eliminar la componente DC de la señal muestreada, s_{in}.

$$s_{of}(n) = s_{in}(n) - s_{in}(n-1) + 0.999 * s_{of}(n-1) \qquad (2.5)$$

2.3.4 Entramado

Como ya hemos dicho, para considerar la señal estacionaria, es necesario procesarla en tramas de corta duración. En el caso de este proyecto, tomamos un periodo de trama de 10 ms.

2.3.5 Cálculo de la medida de la energía

Uno de los coeficientes que se quieren extraer, como se ha dicho anteriormente es el logaritmo de la energía de cada trama.

$$\log E = \ln\left(\sum_i s_{of}(i)^2\right) \qquad (2.6)$$

2.3.6 Pre-énfasis

El filtro de pre-énfasis tiene como objetivo el hecho de que la señal de voz tiene un contenido más elevado en bajas que en altas frecuencias. La siguiente expresión hace que el espectro presente un aspecto más plano:

$$s_{pe}(n) = s_{of}(n) - 0.97 * s_{of}(n-1) \qquad (2.7)$$

2.3.7 Enventanado

El entramado de la señal se puede considerar como la multiplicación de esta por una señal rectangular, lo que en el espacio frecuencial se traduce en convolucionar el espectro de la señal de audio con una sinc. Para evitar en lo posible la aparición de componentes en alta frecuencia, debidas a las discontinuidades de la señal rectangular, se aplica una ventana de Hamming en el proceso de enventanado de la señal.

La expresión de la ventana de Hamming que se utiliza es:

$$s_w(n) = \left\{ 0.54 - 0.46 * \cos\left(\frac{2\pi(n-1)}{N-1} \right) \right\} s_{pe}(n) \qquad (2.8)$$

Vemos en las siguientes imágenes la representación de las ventanas rectangular y de Hamming, en el dominio del tiempo y de la frecuencia.

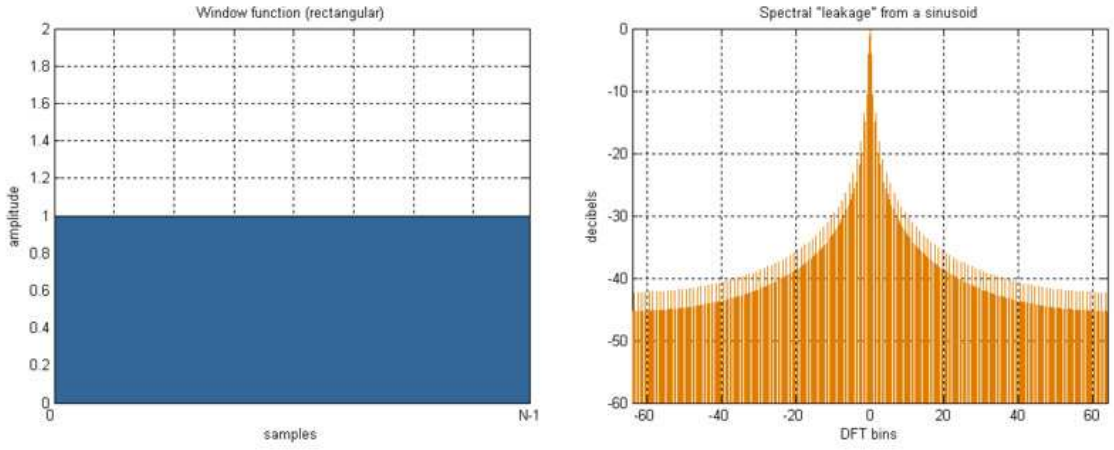

Figura 2.7: *Ventana rectangular*

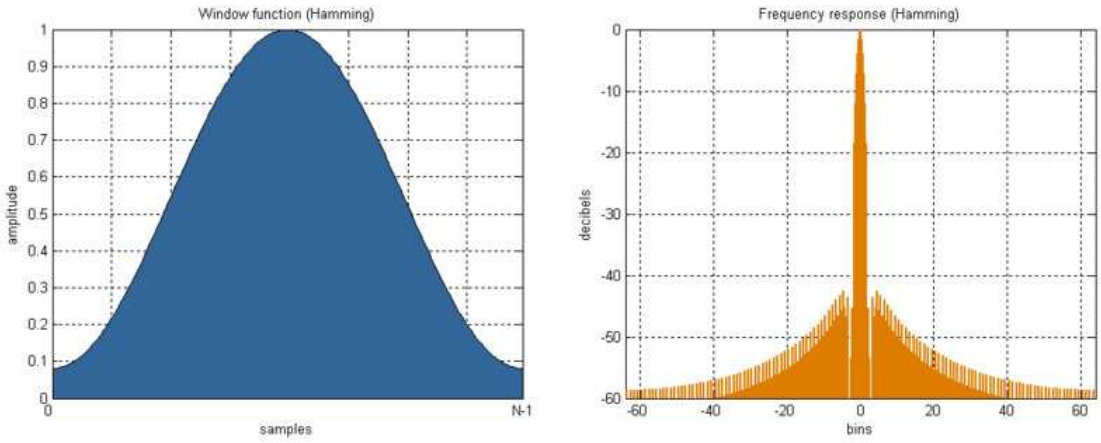

Figura 2.8: *Ventana de Hamming*

Como observamos en las imágenes, el espectro de la ventana de Hamming es más similar a una delta en el dominio frecuencial, por lo que la distorsión que se introduce en el espectro de la señal de audio es menor que en el caso de la ventana rectangular.

2.3.8 Transformada rápida de Fourier

Los segmentos de audio no son más que fluctuaciones de la energía a lo largo del tiempo en diferentes bandas frecuenciales. Para calcular estos parámetros podemos calcular el módulo de la transformada de Fourier a corto plazo (*Short Time Fourier Transform*, STFT). Se obtiene como resultado, una estimación de la densidad espectral de potencia de la señal de audio, habiendo asumido su estacionariedad local.

Para la frecuencia de muestreo del formato PCM (22050 Hz), y la longitud de trama que se toma en este proyecto (20 ms), tenemos 441 muestras por trama. Según el estándar, el tamaño de los arrays para calcular la FFT debe ser potencia de dos, así que completamos ese array con ceros hasta obtener uno de 512 muestras. A partir de esta trama extendida se calcula su FFT.

A continuación se muestra la expresión empleada para calcular la FFT, donde s_w es la trama de entrada, FFTL es la longitud del bloque de FFT y bin_k es el módulo de la transformada de Fourier de salida. Debido a la simetría de la respuesta, únicamente se emplean la mitad de los coeficientes.

$$bin_k = \left| \sum_{n=0}^{FFTL-1} s_w(n) * \exp\left(-j * n * k \frac{2\pi}{FFTL} \right) \right| \qquad k = 0,\ldots,FFTL-1 \qquad (2.9)$$

2.3.9 Filtrado de Mel

Para hacer una aproximación al funcionamiento del oído humano, que no presenta una respuesta lineal en frecuencia, como vemos en la siguiente imagen:

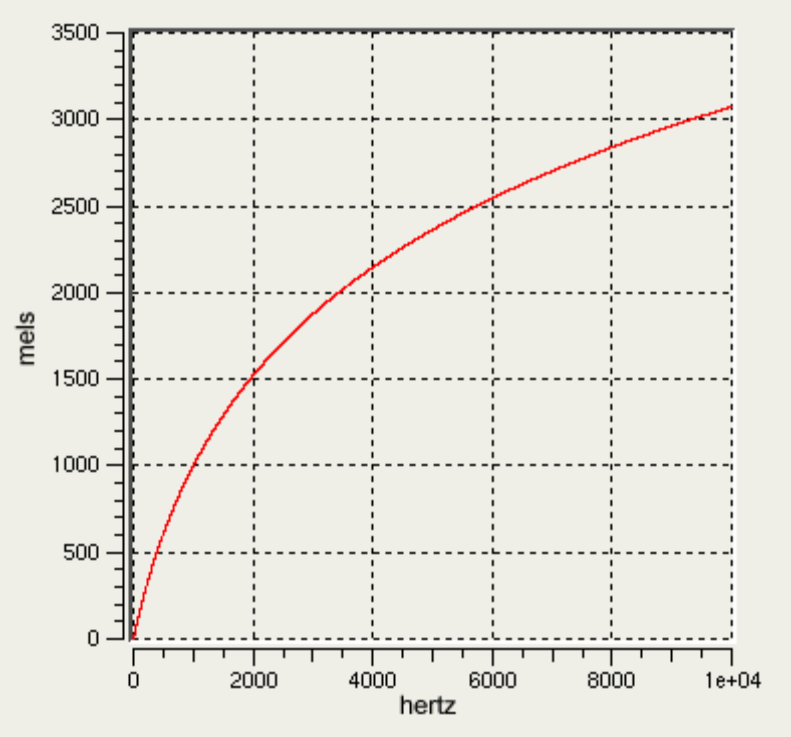

Figura 2.9: *Escala de Mel*

se realiza un filtrado mediante un banco de filtros en escala Mel:

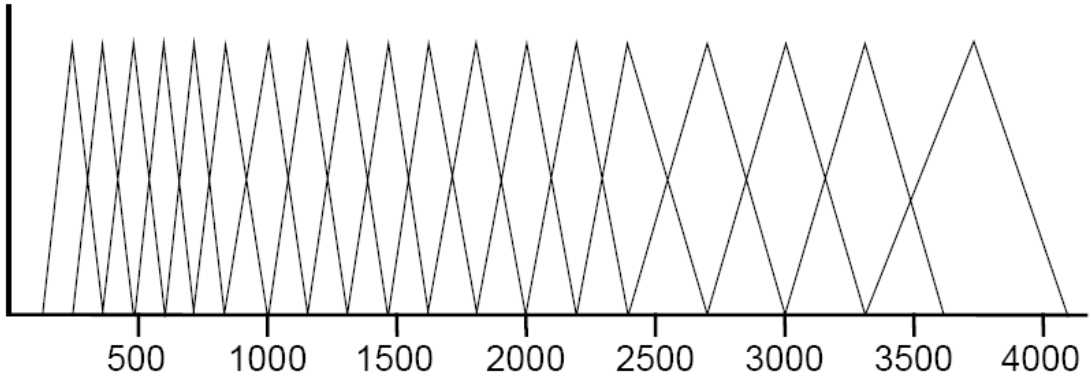

Figura 2.10: *Banco de filtros de Mel*

Para transformar frecuencias expresadas en hertzios a mels, basta con aplicar la siguiente expresión:

$$Mels = 2595 \times \log\left(\frac{1 + f(Hz)}{700}\right) \qquad (2.10)$$

El resultado que se obtiene tras la formación de las bandas de Mel viene dado por:

$$Y(i) = \sum_{i} F_{med}(i,k) * |bin_k| \qquad (2.11)$$

2.3.10 Transformación no lineal y DCT

Las últimas operaciones para extraer los coeficientes MFCC consisten en aplicar el logaritmo neperiano y la transformada de coseno discreta (DCT) a la salida obtenida tras el filtrado de Mel. Las expresiones de estas operaciones, siendo N el número de canales Mel en que se divide la banda de audio, son:

$$f_i = \ln\left[Y(i)\right] \; i = 1,...,N \qquad (2.12)$$

$$C_i = \sum_{j=1}^{N} f_i * \cos\left(\frac{\pi * i}{N}(j - 0.5)\right) \qquad i = 0,...,12 \qquad (2.13)$$

El coeficiente C_0 es redundante si se ha calculado la energía de la trama.

2.3.11 Salida del Front-End

El último paso es calcular la primera derivada de los coeficientes de salida obtenidos, para tener en cuenta la dependencia de cada una de las tramas con la anterior, y así tener una medida de la correlación existente entre ellas. Estos parámetros son los llamados coeficientes delta.

Para extraer estos coeficientes, el método a utilizar se presenta en la ampliación del estándar seguido hasta ahora, el ES 202 050. Para su obtención se utiliza la expresión:

$$d(i,t) = -c(i,t-4) - 0.75 \times c(i,t-3) - 0.5 \times c(i,t-2) - 0.25 \times c(i,t-1) + 0.25 \times c(i,t+1) + 0.5 \times c(i,t+2) + 0.75 \times c(i,t+3) + c(t+4)$$
$$1 \le i \le 12 \qquad (2.14)$$

2.4 Normalización de los parámetros (*Cepstral Mean Normalization*, CMN)

La CMN ha sido habitualmente con MFCC, comúnmente en aplicaciones de reconocimiento del habla. Consiste en un cálculo de la media de los vectores de características a lo largo de la totalidad de datos, y resta el valor medio a cada uno de los vectores cepstrales de ese sonido [12].

Ha demostrado mejorar las tasas de reconocimiento, aliviar los efectos de filtrados lineales o la distorsión convolucional que se introduce en los canales de transmisión o en la grabación.

Sin embargo, no discrimina entre silencio y sonido al computar la media, y no puede ser usada en tiempo real porque requiere del sonido entero para calcular la media cepstral.

3 Técnicas de clasificación de audio

Este capítulo trata de determinar el sistema de clasificación que se usará en este proyecto. Muestra distintas posibilidades comúnmente utilizadas para este fin, y determina cuál de ellas puede servir mejor al fin que se ha propuesto para este trabajo.

3.1 Introducción

El proceso de clasificación de registros de audio consta de dos fases: entrenamiento y evaluación o test. La fase de entrenamiento trata de extraer las características de cada clase acústica, que definen un patrón concreto. La fase de test permite comprobar que el sistema de clasificación discrimina con exactitud los diferentes tipos de archivos, a partir de los patrones representativos de cada uno que se han obtenido en la primera fase de entrenamiento.

Para esta finalidad, en la literatura se han descrito tres técnicas básicas de reconocimiento de patrones que han demostrado ser efectivos en objetivos similares, como reconocimiento de instrumentos, de hablante, etc. Estas técnicas son las máquinas de vectores soporte (*Support Vector Machines*, SVM), las redes neuronales (*Neural Networks*, NN), y los modelos de mezclas de gaussianas (*Gaussian Mixture Models*, GMM).

En este proyecto utilizaremos los GMM, que han demostrado resultados tan buenos o mejores que los otros sistemas y cuyo uso está muy extendido, entre otras, en tareas de procesado de voz. Además, su uso viene también motivado por la capacidad de las mezclas de gaussianas para modelar funciones de densidad de probabilidad arbitrarias.

A continuación se analizan brevemente las diferentes técnicas de clasificación, y por qué se llega a la conclusión de que lo mejor para el objetivo de este proyecto es la opción que se ha tomado.

3.2 Técnicas de clasificación

3.2.1 Redes neuronales

Las redes neuronales tratan de emular la capacidad humana de aprendizaje y aplicación de lo aprendido a nuevas situaciones para tomar decisiones, es decir la memorización y la asociación. [8]

Existen una serie de problemas que no se pueden solucionar mediante un algoritmo, y que el hombre las resuelve mediante la experiencia adquirida en otras situaciones. La solución a estos problemas pasa entonces por la construcción de un sistema que reproduzca esta característica humana.

En resumen, una red neuronal es "un nuevo sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona". Individualmente, las neuronas son un componente relativamente simple del ser humano, pero cuando millares de ellas se conectan conjuntamente se hacen muy poderosas.

Tenemos, entonces, que las redes neuronales:

- Consisten de unidades de procesamiento que intercambian datos o información.

- Se utilizan para reconocer patrones (imágenes, secuencias de tiempo, etc.).

- Tienen capacidad de aprender y mejorar su funcionamiento.

Existen dos tipos de redes neuronales según su similitud con la realidad biológica. Está el modelo de tipo biológico, que trata de simular funciones biológicas como la audición o funciones básicas de la visión. El otro modelo, es el dirigido a aplicación, que está fuertemente ligado a las necesidades de la aplicación para la que se diseña. Este último tipo sería el que sería más adecuado para tareas de clasificación de audio.

Las ventajas que ofrece este sistema son:

- Aprendizaje adaptativo: permite aprender a realizar tareas basadas en un entrenamiento o en una experiencia inicial.
- Auto-organización: las redes neuronales pueden crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.

- Tolerancia a fallos: la destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño.

- Operación en tiempo real: los cálculos neuronales se pueden realizar en paralelo; se diseñan para ello y se fabrican máquinas con hardware especial para obtener esta capacidad.

- Fácil inserción dentro de la tecnología existente: se pueden obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Ello facilitará la integración modular en los sistemas existentes.

Una red neuronal está constituida por neuronas interconectadas y dispuestas en capas. Los datos ingresan por la “capa de entrada”, pasan por la “capa oculta” y salen por la “capa de salida”. La capa oculta puede estar formada por varias capas.

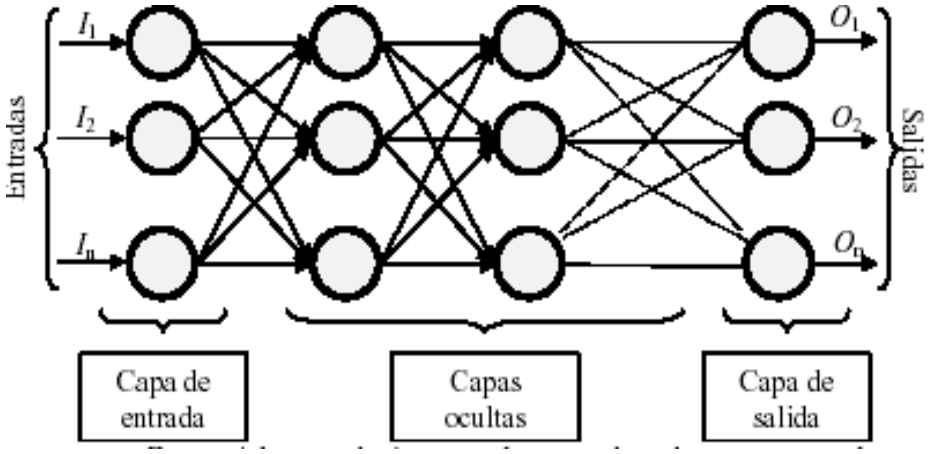

Figura 3.1: *Estructura de capas de una red neuronal*

La neurona artificial pretende mimetizar las características más importantes de las neuronas biológicas. Cada neurona i-ésima está caracterizada en cualquier instante por un valor numérico llamado valor o estado de activación $a_i(t)$; asociado a cada unidad, existe una función de salida, f_i, que transforma el estado actual de activación en una señal de salida. Esta señal se envía a través de los canales de comunicación unidireccionales a otras unidades de la red; en estos canales la señal se modifica de acuerdo con la sinopsis (el peso, w_{ji}) asociada a cada uno de ellos según una determinada regla. Las señales moduladas que han llegado a la unidad j-ésima se combinan entre ellas, generando así la entrada total.

$$Net_j = \sum_i y_i w_{ji} \qquad (3.1)$$

Una función de activación, F, determina el nuevo estado de activación $a_j(t+1)$ de la neurona, teniendo en cuenta la entrada total calculada y el anterior estado de activación $a_j(t)$.

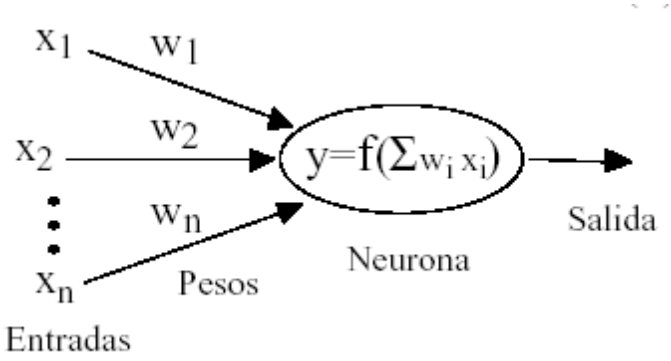

Figura 3.2: *Entorno de una neurona dentro de una red*

3.2.2 Máquinas de Vectores Soporte (SVM)

Las máquinas de soporte vectorial o máquinas de vectores soporte [9, 13] son un conjunto de algoritmos desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T. Pertenecen a la familia de los clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad (introducidos por funciones núcleo o *kernel*) con un sesgo inductivo muy particular (maximización del margen).

Inicialmente se usaron para problemas de clasificación binaria, pero después se ha extendido su uso a problemas de regresión, agrupamiento, multiclasificación, regresión ordinal, y se está trabajando en la resolución de problemas más complejos (árboles y grafos).

Un dato es visto como un punto definido por un vector p-dimensional (una lista de p números), y lo que se quiere saber es cómo separar esos datos con un hiperplano (p-1)-dimensional. Es lo que se llama un clasificador lineal. De estos hiperplanos se quiere obtener el que logra una mayor separación o margen entre las clases de datos. Por lo tanto se quiere maximizar la distancia entre el hiperplano y los datos más cercanos de cada una de las clases; es lo que se llama clasificador de máximo margen.

Si tenemos unos datos de entrenamiento, que son un conjunto de puntos de la forma $\{(x_1,c_1),(x_2,c_2),...,(x_n,c_n)\}$ donde c_i puede ser 1 o -1, según la clase a la que pertenece el punto x_i, se quiere obtener el hiperplano de máximo margen que divida ambas clases.

Ese hiperplano será el conjunto de puntos x que satisface $w \cdot x - b = 0$, siendo w un vector normal, perpendicular al hiperplano. El parámetro b determina la distancia del hiperplano al origen a lo largo del vector normal w.

Para encontrar w y b tal que el margen o distancia entre los hiperplanos paralelos que separan los datos sea el mayor se tiene que cumplir que

$$w \cdot x - b = 1 \qquad (3.2)$$

$$\text{y } w \cdot x - b = -1 \qquad (3.3)$$

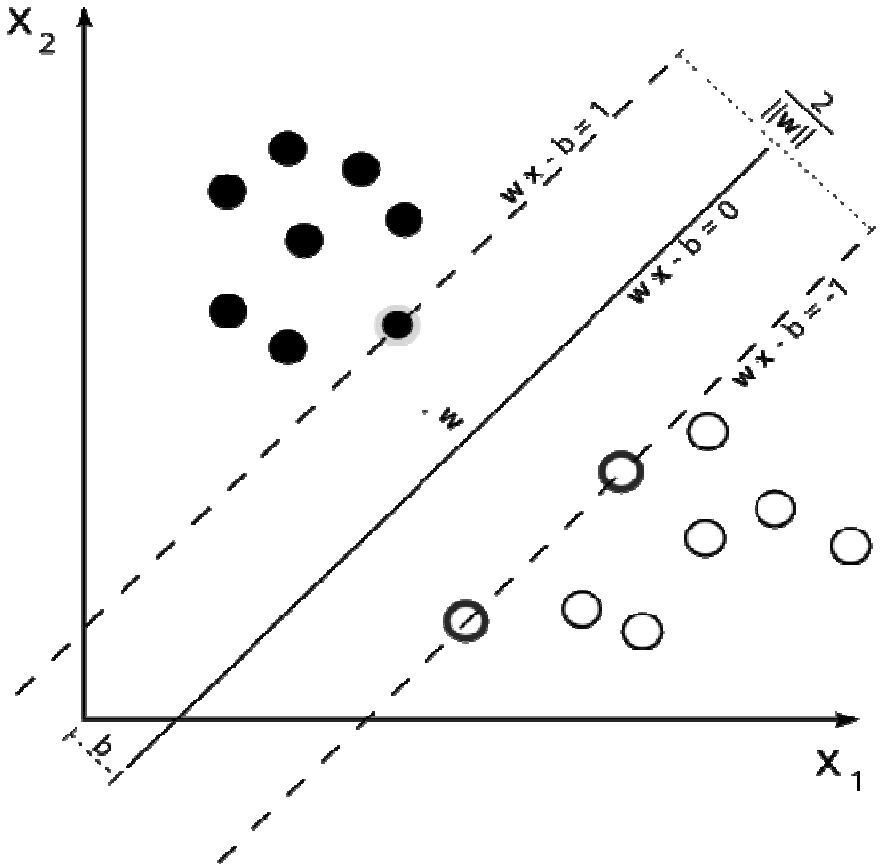

Figura 3.3: *Hiperplano separando dos clases distintas*

Si los datos de entrenamiento son linealmente separables, podemos obtener los hiperplanos del margen de forma que no haya puntos entre ellos, y maximizar la distancia que les separa. Mediante la geometría vemos que la distancia entre ambos planos será $2/|w|$, así que lo queremos es minimizar $|w|$. Teniendo en cuenta los puntos que coinciden con los márgenes tenemos la siguiente expresión para los puntos x_i de la

primera clase: $w \cdot x_i - b \geq 1$ y la siguiente para los de la segunda: $w \cdot x_i - b \leq -1$. Esto puede ser reescrito como:

$$c_i \left(w \cdot x_i - b \right) \geq 1 \text{, para todo } 1 \leq i \leq n \qquad (3.4)$$

Así el problema consiste en hallar w y b, tal que minimicen $|w|$, cumpliendo la anterior expresión.

No siempre es posible separar las clases totalmente con un hiperplano. En este caso hay que minimizar los errores de clasificación; es lo que se ha llamado margen blando. Se introduce ahora ξ_i que es una medida de la clasificación errónea, quedando la siguiente expresión:

$$c_i \left(w \cdot x_i - b \right) \geq 1 - \xi_i \text{ para } 1 \leq i \leq n \qquad (3.5)$$

Se trata de minimizar simultáneamente $|w|$ y ξ_i. La minimización de $|w|$ es equivalente a la de $\frac{1}{2}\|w\|^2$, que es más simple de calcular. Por todo esto tenemos que buscar

$$\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \,, \qquad (3.6)$$

$$\text{tal que } c_i \left(w \cdot x_i - b \right) \geq 1 - \xi_i \text{ para } 1 \leq i \leq n \,. \qquad (3.7)$$

Posteriormente se ha introducido en el esquema de las SVMs el concepto de clasificación no lineal. Para ello, cada producto escalar se sustituye por una función *kernel* (transformación) no lineal. Esto permite hallar el hiperplano de máximo margen en el espacio de características transformado. La transformación puede ser no lineal y el espacio transformado de alta dimensión; así aunque el clasificador es un hiperplano en el espacio de alta dimensión, puede ser no lineal en el espacio original.

3.2.3 Modelos de mezclas de gaussianas (GMM)

El uso de modelos de mezcla de gaussianas (GMM) para tareas de clasificación de audio está motivado por la interpretación de que diferentes componentes gaussianas (o combinaciones lineales de las mimas) sirven para representar diferentes tipos de audio, y por su capacidad para modelar funciones de densidad de probabilidad arbitrarias.

Dicho de otro modo, en primer lugar, las componentes individuales gaussianas en un GMM tienen capacidad para modelar algunas clases acústicas generales y en segundo lugar, se ha mostrado que una densidad de mezclas de gaussianas proporciona una aproximación más fiel a la distribución subyacente de las observaciones obtenidas de sonidos a largo plazo.

Una función de densidad de mezcla de gaussianas es una suma ponderada de *M* componentes con densidades gaussianas, según la ecuación:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \ , \qquad (3.8)$$

donde $\vec{x}$ es un vector aleatorio de dimensión *D*, $b_i(\vec{x}), i = 1,...,M$, son las densidades componentes y $p_i, i = 1,...,M$, son los pesos de las mezclas. Cada densidad componente es una función gaussiana multivariable de *D* componentes, de la forma:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \left| \sum_i \right|^{1/2}} \exp\left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_i)' \sum_i^{-1} (\vec{x} - \vec{\mu}_i) \right\} \qquad (3.9)$$

con vector de media $\vec{\mu}_i$ y matriz de covarianza $\sum_i$. Los pesos de las mezclas satisfacen la restricción $\sum_{i=1}^{M} p_i = 1$. En la figura 3.4 se ve un esquema representativo de la densidad de mezcla de gaussianas con M componentes.

La densidad de mezclas de gaussianas completa se parametriza mediante los vectores de media, matrices de covarianza y pesos de mezclas de todas las densidades componentes. Estos parámetros son representados colectivamente por la notación $\lambda = \{p_i, \vec{\mu}_i, \sum_i\}, i = 1,...,M$. En nuestro caso, cada tipo de sonido (habla, música con voz, y música instrumental) estará representado por un GMM, y referido a su modelo λ .

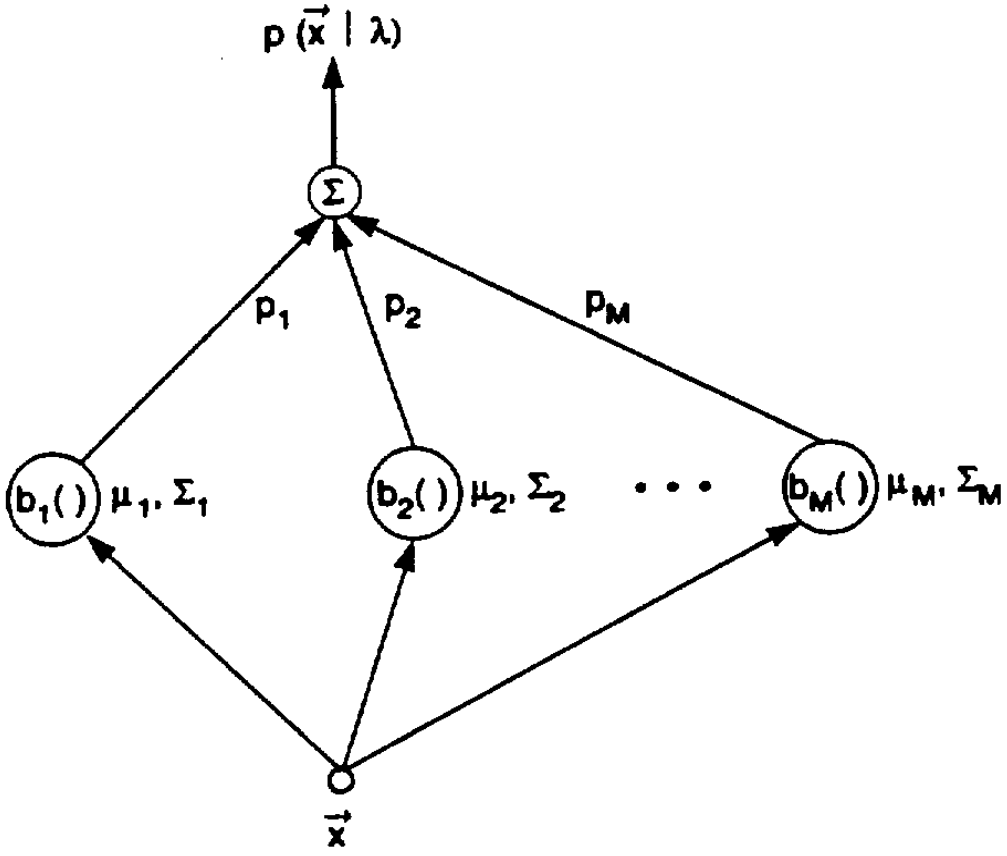

Figura 3.4: *Densidad de mezcla de gaussianas de M componentes*

Los GMM pueden tener varias formas diferentes dependiendo de la elección de las matrices de covarianza. El modelo puede tener una matriz de covarianza por componente gaussiana como en la expresión anterior de λ (covarianza nodal), una matriz de covarianza para todas las componentes gaussianas en un modelo de audio (gran covarianza), o una sola matriz de covarianza compartida por todos los modelos de audio (covarianza global). Las matrices de covarianza pueden también ser completas o

diagonales. Se usan matrices de covarianza nodales y diagonales para modelar los tipos de audio.

El efecto de usar un conjunto de M Gaussianas de covarianza diagonales puede ser igualmente obtenido usando un conjunto mayor de Gaussianas de covarianza diagonales.

Como el entrenamiento y el test de los sonidos están sin etiquetar, las clases acústicas están "escondidas" con lo que la clase de una observación es desconocida. Asumiendo vectores de características independientes, la densidad de observación de vectores de características obtenida de estas clases acústicas escondidas es una mezcla de gaussianas.

Uno de los potentes atributos de los GMM es su habilidad para formar aproximaciones ajustadas a densidades de formas arbitrarias. El modelo gaussiano unimodal clásico representa una distribución de características del audio por una posición (vector de medias) y una forma elíptica (matriz de covarianzas) y el modelo VQ representa una distribución del sonido por un conjunto discreto de plantillas de características. De alguna forma el GMM actúa como un híbrido entre estos dos modelos usando un conjunto discreto de funciones Gaussianas, cada una con su propia media y matriz de covarianza, para permitir una mejor capacidad de modelado. El GMM no sólo proporciona una distribución global ajustada, sus componentes también detallan claramente la naturaleza multimodal de la densidad.

Hay varias técnicas compatibles para estimar los parámetros de un GMM. El método más popular y asentado es la estimación de máxima verosimilitud (*Maximum Likelihood*, ML).

El objetivo de la estimación ML es encontrar los parámetros del modelo que maximizan la verosimilitud de los GMM, dados los datos de entrenamiento. Para una secuencia de T vectores de entrenamiento $X = \{\vec{x}_1, ..., \vec{x}_T\}$, la similitud GMM puede ser escrita así:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda) \qquad (3.10)$$

Desafortunadamente, esta expresión es una función no lineal de los parámetros λ y la maximización directa no es posible. Sin embargo, la estimación de la máxima verosimilitud puede ser obtenerse iterativamente usando un caso especial del algoritmo maximización de la esperanza (*Expectation-Maximization* , EM).

La idea básica del algoritmo EM es, empezando con un modelo inicial λ, estimar un nuevo modelo $\overline{\lambda}$, tal que $p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$. El nuevo modelo llega entonces a ser el modelo inicial para la siguiente iteración y el proceso es repetido hasta que se alcanza algún umbral de convergencia.

En cada iteración del EM, se usan las siguientes fórmulas de reestimación, que garantizan un incremento monótono del valor de verosimilitud del modelo:

Pesos de las mezclas:

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \qquad (3.11)$$

Medias:

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} \qquad (3.12)$$

Varianzas:

$$\overline{\sigma}_i^{\,2} = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} - \overline{\mu}_i^2 \qquad (3.13)$$

donde $\sigma_i^2, x_t,$ y μ_i se refieren a elementos individuales de los vectores $\vec{\sigma}_i^2, \vec{x}_t,$ y $\vec{\mu}_i$, respectivamente.

Las probabilidades a posteriori para la clase acústica i está dada por

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \qquad (3.14)$$

Existen dos factores críticos en el entrenamiento de un modelo de mezclas de gaussianas para cada tipo de audio: en primer lugar, la selección del número de mezclas M y en segundo lugar, la inicialización de los parámetros de los modelos previos a la aplicación del algoritmo EM. No hay buenos medios teóricos para guiar en estas selecciones, así que suelen determinarse experimentalmente para una tarea dada.

En el proceso de clasificación del tipo de audio en sí, se considera un grupo de S de clases acústicas de audio $S = \{1, 2, ..., S\}$ representados por sus GMMs respectivos $\lambda_1, \lambda_2, ..., \lambda_S$. Dichos modelos han sido obtenidos en la fase de entrenamiento explicada anteriormente. El objetivo es encontrar el modelo de sonido que tiene la máxima probabilidad a posteriori para una secuencia de observación dada. Formalmente,

$$\hat{S} = \arg\max_{1 \le k \le S} \Pr(\lambda_k \mid X) = \arg\max_{1 \le k \le S} \frac{p(X \mid \lambda_k) \Pr(\lambda_k)}{p(X)} \qquad (3.15)$$

donde la segunda ecuación se obtiene aplicando la regla de Bayes. Suponiendo que todas las clases acústicas son equiprobables (i.e., $\Pr(\lambda_k) = 1/S$) y que $p(X)$ es la misma para todos los modelos de sonido, la regla de clasificación se simplifica en:

$$\hat{S} = \arg\max_{1 \le k \le S} p(X \mid \lambda_k) \qquad (3.16)$$

Usando logaritmos y suponiendo que las observaciones son independientes entre sí, el sistema de clasificación determina la clase acústica de audio a través de la siguiente expresión:

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{t=1}^{T} \log p\left(\vec{x}_t \mid \lambda_k\right) \qquad (3.17)$$

en el que $p\left(\vec{x}_t \mid \lambda_k\right) = \sum_{i=1}^{M} p_i b_i\left(\vec{x}\right)$ (3.18)

4 Marco experimental

En este capítulo se define el entorno en que se han realizado los experimentos, la base de datos, el sistema de clasificación de registros de audio diseñado y las herramientas utilizadas. También se especifica en qué consiste cada uno de los experimentos que se realizarán, y que servirán para determinar qué parámetros de la señal de audio son más válidos para la correcta diferenciación de los diferentes tipos de señales de audio (habla, música instrumental, y música con voz).

4.1 Base de datos

La base de datos inicial está compuesta por 244 archivos de audio: 80 archivos de música instrumental, 84 que contienen música con voz, y otros 80 que contienen habla, sin música. La duración de cada uno de estos archivos es de 15 segundos y su frecuencia de muestreo es de 22050 Hz.

Dado que el número de ficheros de la base de datos es reducido y sin embargo, la duración de cada uno de dichos ficheros es bastante elevada, se decidió proceder a la división de cada uno de los archivos en 10 fragmentos de igual duración (1.5 segundos). De este modo, se disponen de 2440 archivos de audio para la experimentación.

El proceso de división se llevó a cabo mediante un programa realizado en MATLAB. A continuación, hubo que realizar un proceso de comprobación manual de los ficheros resultantes puesto que al dividir los archivos de música con voz, muchos siguieron manteniendo este estatus, pero otros perdieron los segmentos con presencia de voz, convirtiéndose en archivos de música instrumental. Por ello hubo escuchar todos estos

archivos para determinar a qué grupo pertenecían definitivamente y reetiquetarlos en caso de que fuera necesario.

Después de todo este proceso queda una base de datos muy diferente de la anterior, mucho más apta para los experimentos que se querían realizar. En resumen, la nueva base de datos consta de 901 archivos de música instrumental, 739 archivos de música con voz, y 800 archivos de habla, que todos juntos conforman los 2440 archivos de segundo y medio de duración que se ha comentado antes.
Finalmente, la base de datos se ha completado con la generación de los ficheros de etiquetas que indican el tipo de registro de audio contenido en cada fichero.

La nueva base de datos sigue siendo reducida, por lo que, en estas condiciones resultará difícil extraer conclusiones lo suficientemente solventes de los experimentos. Para solucionar este problema, para los experimentos se utilizó el método *leave-one-out*. Básicamente, consiste en dividir la base de datos en N grupos (en nuestro caso, seis). Se entrena el sistema con los archivos contenidos en N-1 de esos grupos, y se realiza el test con los archivos del grupo restante. La operación se repite N veces, cambiando en cada ocasión el grupo de test, de manera que este (y por tanto el conjunto de entrenamiento también) sea diferente en cada una de las ejecuciones. Finalmente, el resultado final se obtiene como promedio de los resultados parciales. Esta técnica proporciona una mayor consistencia a las pruebas y una mayor fiabilidad a los resultados obtenidos.

4.2 Sistema de clasificación de registros de audio

El sistema de clasificación utilizado en este proyecto está basado en modelos de mezcla de gaussianas (GMM) descritos en el capítulo 3 (apartado 3.2.3). Se consideran tres modelos para clasificar, que corresponden con los tres tipos de audio que se tratan: habla, música con voz y música instrumental.

En la figura 4.1. se muestra el esquema del sistema global, que se compone básicamente del módulo de parametrización y del de clasificación. El módulo de parametrización es el que se ha descrito en el capítulo 2, y sobre el que se trabaja en este proyecto,

buscando la mejor caracterización posible para los archivos de audio. El módulo de clasificación se compone del reconocedor en sí, y de los modelos que se generan mediante los GMM, para cada una de las clases de audio recogidas en el diccionario.

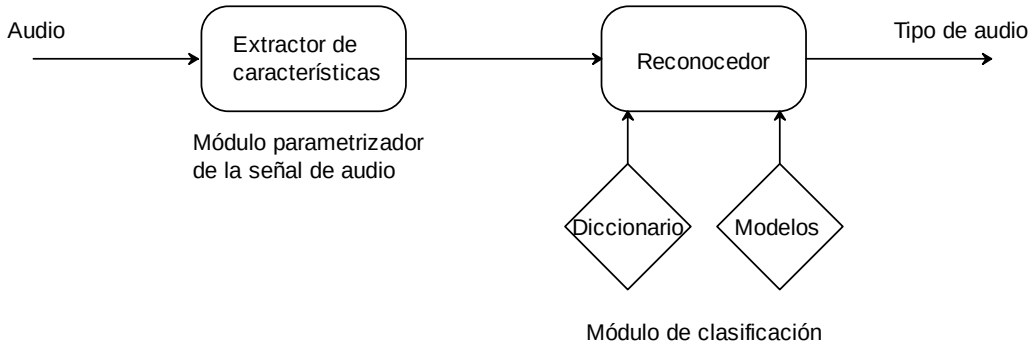

Figura 4.1: *Esquema de bloques del sistema global*

4.2.1 Módulo de parametrización

En cuanto a la extracción de parámetros, la intención de este trabajo es experimentar con parámetros MFCC, ASE, y ASP, agregándoles en cada uno de los experimentos, en algunos casos, alguna característica adicional.

Este módulo es el fundamental del proyecto, ya que el objetivo de este es configurar este módulo de manera que a la salida tengamos las mejores tasas de reconocimiento. Se encarga de convertir las señales de audio en un conjunto de datos que representan a la señal anterior. Lo que se busca es conseguir el mejor conjunto de datos posible. [7]

Los primeros parámetros con los que se prueba son los MFCC (Mel Frequency Cepstral Coefficients), que ya han sido habitualmente utilizados en otras aplicaciones para la parametrización de la señal de voz. Se comprobará en este proyecto qué tal se adaptan a la finalidad de éste.

Para la extracción de los parámetros MFCC, a la señal ,dividida en tramas para su parametrización, se le aplica un pre-énfasis de factor 0.97. Para la división en tramas se

utiliza una ventana de Hamming de 25 ms de duración, y un desplazamiento entre ellas (por lo tanto, también de tramas) de 10 ms. Se aplica la transformada rápida de Fourier (FFT) para obtener el espectro de cada trama, se calcula la log-energía en cada banda mediante un banco de 40 filtros en escala de Mel, y se les aplica la transformada de coseno discreta (DCT), y se obtienen los 12 primeros MFCC.

Para la extracción de los parámetros ASE se divide la señal de audio en tramas solapadas, se configura el periodo de trama como 10 ms, y el tamaño de ventana de Hamming como 25ms. Después se analiza mediante una transformada de Fourier a corto plazo (STFT). Después, los coeficientes espectrales se agrupan en sub-bandas logarítmicas para reducir la dimensionalidad. Los canales se espacian según una resolución de octava que variará en los experimentos, y que van desde un límite inferior de 62.5 Hz a un límite superior de 11025 Hz (la mitad de la frecuencia de muestreo).

Para la extracción de los parámetros MFCC se utilizó la herramienta HCopy de HTK, que se encarga automáticamente de ello, a partir de las configuraciones que se han comentado previamente.

En los experimentos que utilizan ASE y ASP, estos parámetros se extraen mediante la herramienta "MPEG-7 Audio Reference Software Toolkit" implementada en MATLAB. Posteriormente, los ficheros de parámetros correspondientes se convierten a formato HTK, para que se puedan utilizar en los módulos de entrenamiento y test implementados con HTK.

4.2.2 Módulo de reconocimiento o clasificación

En la fase de entrenamiento se crean los modelos que posteriormente serán utilizados para la clasificación del tipo de audio. Se utiliza el método de reestimación descrito en el capítulo anterior. Para cada tipo de audio, en las diferentes fases del proceso, se generan modelos GMM con diferente número de de mezclas (1, 2, 4, 8, 16 y 32), con el objeto de determinar que número de mezclas gaussianas es el más apropiado a partir de los resultados de los tests.

En la fase de reconocimiento o test, se calcula la log-verosimilitud del vector de observaciones acústicas de entrada con respecto a cada uno de los modelos acústicos considerados. Finamente, se asigna aquella clase que produce la mayor log-verosimilitud. La gramática del sistema es muy simple, ya que se considera que en cada fichero de audio sólo hay una única clase acústica (habla, o música con voz, o bien música instrumental). [4, 10]

En este trabajo, todo el proceso se realiza a través de la herramienta HTK, cuyas componentes son archivos ejecutables desde MS-DOS. Para el entrenamiento se utiliza el programa HERest, y para la clasificación el programa HVite. El programa HResults proporciona los resultados que nos permiten interpretar qué tal funciona cada uno de los experimentos.

Todos los experimentos se han automatizado en la medida de lo posible mediante la creación de una serie de listas de ficheros, ficheros de configuración y scripts configurables.

4.3 Protocolo experimental

Los experimentos que se han llevado a cabo en este proyecto tienen, como ya se ha mencionado antes, como objetivo determinar qué parámetros o características que definen los archivos de audio permiten una mejor diferenciación entre las tres clases que se han propuesto. Por ello se han realizado diversos experimentos con diferentes parámetros acústicos:

- Parámetros MFCC
- Parámetros MFCC con log-energía
- Parámetros MFCC con parámetros delta (primera derivada de los anteriores)
- Parámetros MFCC con log-energía y parámetros delta
- Parámetros MFCC con CMN
- Parámetros MFCC con log-energía y CMN
- Parámetros MFCC con parámetros delta y CMN
- Parámetros MFCC con log-energía, parámetros delta y CMN

- Parámetros ASE con resolución de octava 1/4, 1/6, y 1/8
- Parámetros ASE con mejor resolución y log-energía
- Parámetros ASE con mejor resolución y deltas
- Parámetros ASE con mejor resolución y parámetros delta y log-energía
- La mejor configuración con los parámetros ASE y CMN
- Parámetros ASP, a partir de los ASE con mejor resolución y considerando diferentes números de bases
- Parámetros ASP con el mejor número de bases y parámetros delta
- Parámetros ASP con el mejor número de bases y parámetros delta y log-energía
- La mejor configuración con los parámetros ASP y CMN

En el capítulo siguiente se describen estos experimentos y los resultados obtenidos.

5 Resultados experimentales

En este capítulo se muestran los resultados obtenidos de la realización de los experimentos propuestos en el capítulo anterior. A la vista de estos resultados se podrán comparar los que se obtienen con unos parámetros y con otros, y de el análisis de éstos se podrán sacar conclusiones sobre cuales de esos parámetros son más apropiados para una buena clasificación de los archivos de audio, que es la finalidad última de este proyecto.

5.1 Introducción

Los módulos de extracción de características que se quieren probar en este proyecto se dividen en tres grupos: el primero, está basado en parámetros MFCC, el segundo en parámetros ASE, y el último en los parámetros ASP. A cada uno de estos conjuntos de coeficientes básicos de la señal de audio se le añadirá alguna característica más en cada uno de los experimentos (log-energía, parámetros de primera derivada, …), para comprobar si mejoran los resultados de clasificación.

Los parámetros MFCC han sido utilizando habitualmente en múltiples aplicaciones de voz, como reconocimiento de habla, o de hablante. Los parámetros ASE y los ASP forman parte del estándar MPEG-7, de más reciente utilización.

Los resultados experimentales vienen dados por el porcentaje de ficheros de audio correctamente clasificados para cada clase acústica y el promedio de estas tasas de reconocimiento calculado sobre todas las clases. De esos resultados se extraerán las conclusiones sobre qué características del audio sirven obtener la mejor clasificación de las tres clases propuestas: habla, música instrumental y música con voz.

5.2 Resultados con parámetros MFCC

En este apartado se muestran los resultados de cada uno de los experimentos realizados con los parámetros MFCC. Se muestran los resultados en tablas que indican el porcentaje de acierto en la clasificación de los archivos para cada clase acústica y el promedio para todas las clases y considerando modelos GMM con diferente número de gaussianas en la mezcla.

5.2.1 Experimento 1

Para este experimento se utilizan únicamente parámetros MFCC, en concreto 12. Los resultados obtenidos con estos parámetros son:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**	**64 mezclas**
Habla	87.90%	91.40%	93.20%	94.00%	95.20%	95.60%	95.50%
Música instrumental	53.60%	60.80%	57.90%	59.40%	60.20%	58.20%	54.90%
Música con voz	62.40%	63.60%	67.40%	66.60%	65.50%	66.40%	69.10%
Total	67.50%	71.68%	72.38%	72.91%	73.28%	72.95%	72.54%

Tabla 5.1: *Experimento 1 (MFCC)*

5.2.2 Experimento 2

Aquí se prueba si el añadir la log-energía a los MFCC mejora la caracterización, y con ello la clasificación. Se obtienen estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**	**64 mezclas**
Habla	96.40%	97.20%	97.10%	97.10%	97.50%	97.40%	97.10%
Música instrumental	55.60%	61.70%	59.20%	57.80%	57.80%	55.40%	54.20%
Música con voz	64.00%	64.50%	67.00%	68.20%	67.00%	66.40%	69.30%
Total	71.52%	74.22%	73.98%	73.85%	73.61%	72.50%	72.83%

Tabla 5.2: *Experimento 2 (MFCC+logE)*

Vemos que la log-energía mejora los resultados para pocas mezclas. Según se aumenta el número de éstas los resultados varían mucho menos respecto al experimento anterior.

5.2.3 Experimento 3

En este experimento se añaden a los parámetros MFCC los parámetros delta, que son la primera derivada de los otros. Se logran los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**	**64 mezclas**
Habla	96.10%	93.90%	89.50%	96.40%	96.80%	96.90%	97.00%
Música instrumental	62.50%	56.90%	60.60%	57.60%	61.00%	60.20%	58.60%
Música con voz	61.70%	76.00%	72.50%	73.90%	71.30%	71.90%	74.60%
Total	73.28%	74.84%	73.69%	75.25%	75.86%	75.74%	76.02%

Tabla 5.3: *Experimento 3 (MFCC+delta)*

Al añadir los parámetros delta sí que se nota una mejora en los resultados para cualquier cantidad de mezclas, respecto del primer experimento. Esto se debe a la mejora de la clasificación en las dos clases de audio musicales, especialmente en la música con voz.

5.2.4 Experimento 4

Este experimento une los parámetros MFCC con la log-energía y los parámetros delta de ambos. Se logran los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**	**64 mezclas**
Habla	97.90%	97.50%	97.10%	98.00%	98.00%	98.00%	97.90%
Música instrumental	61.50%	57.50%	62.90%	61.30%	62.70%	57.40%	55.70%
Música con voz	66.20%	76.20%	69.60%	72.30%	72.00%	74.20%	72.90%
Total	74.84%	76.27%	76.15%	76.64%	77.09%	75.78%	74.75%

Tabla 5.4: *Experimento 4 (MFCC+logE+delta)*

La combinación de estas tres características ofrece una mejora notable hasta 16 mezclas, de ahí en adelante, los resultados son peores que en el experimento 3.

5.2.5 Experimento 5

En este experimento se le aplica a los 12 parámetros MFCC que se utilizaban en el experimento 1 la normalización CMN. Se obtienen los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**	**64 mezclas**
Habla	97.00%	97.20%	96.60%	96.90%	97.00%	97.10%	97.20%
Música instrumental	77.00%	78.60%	78.50%	79.00%	80.10%	80.80%	82.10%
Música con voz	46.30%	49.10%	57.00%	56.80%	57.20%	59.50%	62.50%
Total	74.26%	75.78%	77.91%	78.16%	78.73%	79.71%	81.15%

Tabla 5.5: *Experimento 5 (MFCC+CMN)*

En este experimento la mejora respecto de los anteriores es muy notable, para toda cantidad de mezclas, pero especialmente en las más altas. Esta mejora se debe, principalmente a la sensible mejora de los resultados para música instrumental, que hasta ahora era la que peores resultados estaba mostrando. A costa de mejorar para esa clase empeora notablemente la clasificación de música con voz. Pero el resultado global es positivo al aplicar la normalización CMN.

5.2.6 Experimento 6

A los parámetros del experimento 2 (MFCC y log-energía), se les aplica ahora la normalización CMN, obteniendo como resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**	**64 mezclas**
Habla	98.10%	98.40%	98.20%	97.90%	97.60%	97.40%	97.90%
Música instrumental	76.80%	73.40%	78.00%	78.70%	80.70%	80.50%	80.90%
Música con voz	50.70%	64.10%	65.50%	67.40%	66.20%	66.40%	67.40%
Total	75.90%	78.77%	80.86%	81.56%	81.84%	81.76%	82.38%

Tabla 5.6: *Experimento 6 (MFCC+logE+CMN)*

La mejora respecto del experimento 2, al aplicar la CMN es muy notable, parece confirmar que la normalización es muy positiva.

5.2.7 Experimento 7

Para este experimento se les aplica a los parámetros del experimento con MFCC y deltas la normalización CMN, obteniéndose los resultados que se muestran a continuación:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**	**64 mezclas**
Habla	97.40%	94.90%	96.40%	97.00%	97.00%	97.50%	97.40%
Música instrumental	74.60%	73.60%	76.50%	78.60%	81.40%	82.00%	81.00%
Música con voz	48.20%	59.30%	59.30%	58.60%	60.60%	62.50%	66.00%
Total	74.06%	76.23%	77.79%	78.57%	80.20%	81.19%	81.84%

Tabla 5.7: *Experimento 7 (MFCC+delta+CMN)*

En este experimento vemos que, efectivamente, la normalización CMN mejora los resultados respecto del mismo experimento sin ella (en este caso ese experimento era el 3). Lo que parece constatarse también es que la mejora más notable se produce en la clasificación de música instrumental.

5.2.8 Experimento 8

En este experimento se le aplica la normalización CMN a los parámetros utilizados en el 4 (MFCC, log-energía, y deltas) teniendo los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**	**64 mezclas**
Habla	98.10%	97.00%	98.00%	98.00%	98.10%	98.00%	98.00%
Música instrumental	76.10%	72.80%	76.80%	75.20%	77.60%	78.50%	79.50%
Música con voz	51.20%	66.30%	64.70%	69.00%	70.50%	71.00%	70.90%
Total	75.78%	78.77%	80.08%	80.82%	82.17%	82.62%	82.95%

Tabla 5.8: *Experimento 8 (MFCC+logE+delta+CMN)*

Nuevamente se ve en la tabla la mejora que produce la normalización CMN, por lo que parece ser una buena idea su aplicación. Mejora la clasificación de todas las clases de audio, pero la que más lo nota es la de música instrumental, para la que resulta especialmente positiva.

5.2.9. Comparativa de los experimentos con MFCC

La primera comparación que se establece es entre la cantidad de mezclas en los GMM, para determinar, hasta qué cantidad de mezclas compensa elevar por su eficiencia éstos. También hay que mantener un compromiso con la complejidad computacional que supone utilizar más mezclas. Por ello se utilizarán las que produzcan una tasa de reconocimiento elevada, y se vea que a partir de ese número el aumento de la eficiencia no es lo suficientemente significativo.

A continuación mostramos una gráfica que representa la evolución del porcentaje total de acierto en la clasificación, frente al número de mezclas utilizadas en cada caso. Se muestran todos los experimentos juntos.

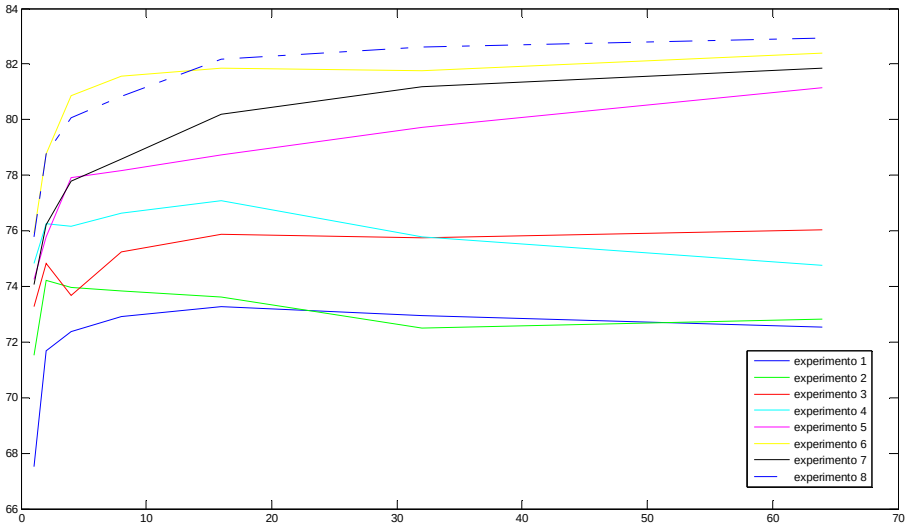

Figura 5.1: *Evolución de la tasa de reconocimiento global frente al número de mezclas, en los 8 primeros experimentos (los que utilizan parámetros MFCC)*

En la figura 5.1 se ve muy claro que no merece la pena aumentar el número de mezclas por encima de 32, ya que para 64 la mejoría en la tasa de reconocimiento o es muy leve, o incluso, como en el primer experimento, desciende. Para algunos casos es incluso mejor 16 mezclas que 32, pero al haber una variación más notable, tanto positiva, como, en algunos casos, negativa, se experimenta también con 32 mezclas.

De la figura 5.1 se puede sacar como conclusión, también que los dos experimentos que mejores resultados arrojan son el 6 (MFCC con log-energía y CMN) y el 8 (MFCC con log-energía, deltas y CMN). El experimento 8 es el que llega a los porcentajes de acierto global más elevados, pero el experimento 6 logra valores cercanos con muchas menos mezclas, ya con 16 mezclas da un valor de acierto que es solamente inferior al obtenido en el experimento 8 con 64 mezclas en poco más de un punto porcentual, como se puede ver en las tablas 5.6 y 5.8.

Otra conclusión importante que puede extraerse de los experimentos es que la aplicación de una técnica de normalización de los parámetros cepstrales (como CMN) incrementa notablemente la tasa de reconocimiento del sistema.

5.3 Resultados con parámetros ASE

En este apartado se presentan los experimentos con los parámetros ASE obtenidos con diferentes resoluciones y con modelos acústicos de 1, 2, 4, 8 16 y 32 gaussianas.

5.3.1 Experimento 9

En este experimento se utilizan parámetros ASE con resolución de octava 1/4. A partir de este experimento se desechan las pruebas con 64 mezclas, ya que no suponen una mejora relevante y representan un importante esfuerzo computacional. Desde ahora se prueba sólo hasta 32 mezclas. Este experimento da los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	74.40%	82.20%	89.20%	88.00%	89.10%	91.80%
Música instrumental	25.60%	23.40%	33.50%	40.40%	49.30%	48.70%
Música con voz	38.20%	56.20%	47.80%	47.80%	50.20%	54.30%
Total	45.41%	52.62%	56.11%	58.24%	62.62%	64.51%

Tabla 5.9: *Experimento 9 (ASE res. octava 1/4)*

5.3.2 Experimento 10

Ahora se experimenta con los parámetros ASE, pero para una resolución de octava de 1/6, obteniendo estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	52.10%	77.20%	88.60%	91.80%	91.80%	93.10%
Música instrumental	57.00%	27.50%	33.20%	44.60%	46.50%	48.40%
Música con voz	36.30%	52.90%	52.60%	47.10%	54.00%	52.20%
Total	49.14%	51.52%	57.25%	60.82%	63.61%	64.22%

Tabla 5.10: *Experimento 10 (ASE res. octava 1/6)*

5.3.3 Experimento 11

En este experimento probamos los parámetros ASE, pero con una resolución de octava de 1/8. Se obtienen los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	55.80%	78.40%	87.20%	91.80%	91.90%	93.10%
Música instrumental	51.40%	25.60%	34.00%	42.20%	47.20%	48.20%
Música con voz	36.50%	53.00%	49.70%	45.70%	53.70%	57.50%
Total	48.32%	51.23%	56.19%	59.51%	63.81%	65.74%

Tabla 5.11: *Experimento 11 (ASE res. octava 1/8)*

5.3.4 Experimento 12

Para los siguientes experimentos se utiliza la resolución de octava para los parámetros ASE que mejor resultado haya dado, y que a la vista de los 3 últimos experimentos, y por no mucha diferencia, parece ser la resolución 1/8. Así pues a partir de este experimento los parámetros ASE tendrán siempre esta resolución de octava. En esta prueba, en concreto, le añadimos a esos parámetros ASE la log-energía, obteniendo estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	59.90%	80.80%	90.20%	93.00%	92.60%	93.60%
Música instrumental	41.20%	23.80%	32.70%	45.40%	46.20%	47.40%
Música con voz	36.90%	56.00%	55.20%	51.20%	55.20%	54.70%
Total	46.02%	52.21%	58.40%	62.75%	64.14%	64.75%

Tabla 5.12: *Experimento 12 (ASE+logE)*

En este caso la inclusión de la log-energía no produce mejoras en todos los casos. Para el caso de 2, 4, 8 y 16 mezclas, los resultados son superiores, pero sin embargo, el funcionamiento del sistema se degrada para 1 y 32 mezclas.

5.3.5 Experimento 13

En este experimento a los parámetros ASE con resolución 1/8 se le suman sus parámetros delta, lo que proporciona estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	47.20%	75.40%	84.50%	89.40%	91.80%	93.40%
Música instrumental	74.70%	54.20%	54.30%	47.30%	55.20%	56.20%
Música con voz	35.60%	51.40%	50.20%	49.50%	58.20%	54.30%
Total	53.85%	60.29%	62.95%	61.76%	68.07%	67.79%

Tabla 5.13: *Experimento 13 (ASE+delta)*

5.3.6 Experimento 14

Para este experimento, además de los parámetros ASE la resolución 1/8, utilizamos la log-energía y los parámetros delta de estas características. Se tienen los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	49.80%	78.00%	88.20%	94.20%	95.50%	96.00%
Música instrumental	73.60%	51.70%	51.50%	49.80%	55.60%	59.00%
Música con voz	36.10%	54.80%	53.50%	50.10%	59.50%	58.10%
Total	54.43%	61.27%	64.14%	64.47%	69.88%	70.86%

Tabla 5.14: *Experimento 14 (ASE+logE+delta)*

5.3.7 Experimento 15

Este experimento consiste en tomar la combinación que mejores resultados arroja con parámetros ASE. En este caso, la mejor combinación es utilizar parámetros ASE con log-energía y añadirle los parámetros delta de todos ellos. A estos parámetros se les aplica, ahora, la normalización CMN, obteniéndose estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	47.90%	65.80%	75.00%	87.40%	87.90%	90.00%
Música instrumental	75.20%	55.50%	52.90%	52.90%	60.50%	61.30%
Música con voz	35.30%	50.60%	49.70%	50.50%	51.00%	55.20%
Total	54.18%	57.38%	59.18%	63.48%	66.60%	68.85%

Tabla 5.15: *Experimento 15 (ASE+logE+delta+CMN)*

5.3.7. Comparativa de los experimentos con ASE

La siguiente comparación que se podría establecer sería entre los experimentos realizados con parámetros ASE, pero diferente resolución de octava, para comprobar, cuál de estas es mejor. En la siguiente figura se comparan las diferentes resoluciones de octava, y se puede ver que con la de 1/8 se logran los mejores resultados, ya desde las 16 mezclas, así que se toma esta como la mejor resolución de octava.

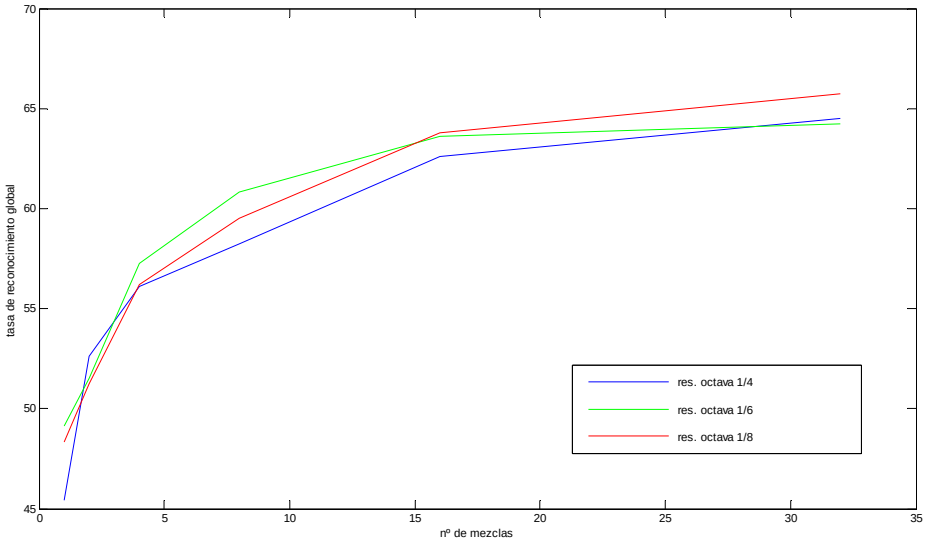

Figura 5.2: *Evolución de la tasa de reconocimiento global frente al número de mezclas, en los experimentos 9, 10 y 11*

Ya con la resolución de octava elegida, se procede a comparar el resto de experimentos realizados con parámetros ASE.

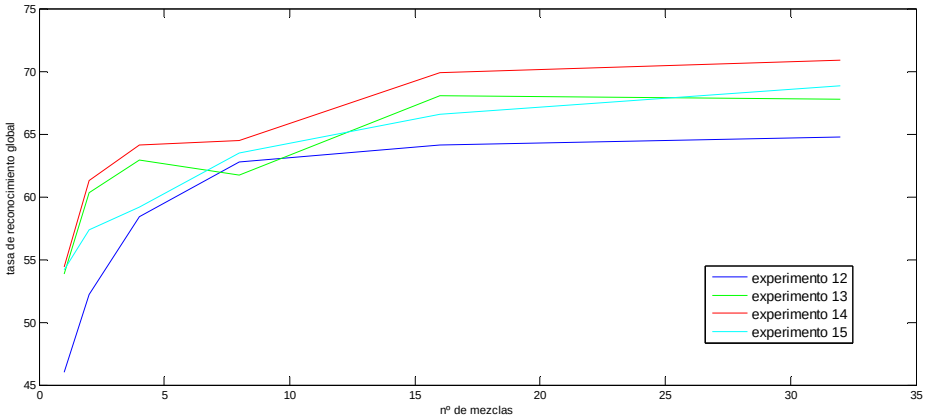

Figura 5.3: *Evolución de la tasa de reconocimiento global frente al número de mezclas, en los experimentos 12, 13, 14 y 15*

A la vista de la figura 5.3 se percibe claramente que el experimento que mejores resultados proporciona para cualquier cantidad de mezclas es el experimento 14, que es el que combinaba parámetros ASE, log-energía, y los parámetros delta de éstos. En este

caso, la aplicación de normalización de los parámetros por la media degrada los resultados del sistema.

5.4 Resultados con parámetros ASP

En este apartado se muestran los experimentos realizados con los parámetros ASP, que parten de los ASE de resolución de octava 1/8. En este caso, además de presentar los resultados con modelos acústicos de diferente número de gaussianas, se varía el número de bases de proyección de los parámetros ASP.

5.4.1 Experimento 16

A partir de este experimento comienzan las pruebas de los parámetros ASP, que parten de los ASE de resolución de octava 1/8. En este primero se prueba con parámetros ASP calculados con 49 bases de proyección, con lo que se obtienen los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	81.20%	95.00%	95.80%	96.40%	97.00%	97.40%
Música instrumental	36.20%	43.00%	52.30%	55.30%	52.60%	56.40%
Música con voz	77.50%	65.20%	57.60%	57.40%	60.40%	60.60%
Total	63.48%	66.76%	68.16%	69.39%	69.51%	71.11%

Tabla 5.16: *Experimento 16 (ASP 49 bases)*

5.4.2 Experimento 17

Para esta prueba se utilizan parámetros ASP calculados a partir de 45 bases de proyección, con estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	84.10%	95.50%	96.80%	96.60%	98.10%	98.10%
Música instrumental	37.70%	42.00%	53.70%	53.80%	55.70%	58.50%
Música con voz	77.70%	72.40%	69.60%	67.80%	68.70%	68.30%
Total	65.04%	68.73%	72.62%	72.09%	73.57%	74.47%

Tabla 5.17: *Experimento 17 (ASP 45 bases)*

5.4.3 Experimento 18

Para calcular los parámetros ASP utilizados en este experimento se emplean 41 bases de proyección, dando estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	84.80%	95.60%	96.20%	97.40%	98.10%	98.20%
Música instrumental	38.10%	43.10%	51.20%	53.40%	55.70%	56.20%
Música con voz	77.30%	72.10%	72.50%	68.10%	69.70%	70.40%
Total	65.25%	69.10%	72.42%	72.25%	73.85%	74.26%

Tabla 5.18: *Experimento 18 (ASP 41 bases)*

5.4.4 Experimento 19

En este experimento se prueban los parámetros ASP extraídos con 37 bases de proyección, y se obtienen estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	83.40%	93.60%	96.10%	97.00%	96.90%	97.00%
Música instrumental	38.30%	45.00%	52.80%	55.50%	55.60%	53.40%
Música con voz	77.00%	60.20%	59.30%	56.60%	58.90%	63.70%
Total	64.80%	65.53%	68.98%	69.43%	70.12%	70.82%

Tabla 5.19: *Experimento 19 (ASP 37 bases)*

5.4.5 Experimento 20

Aquí se experimenta con 33 bases de proyección para extraer los parámetros ASP. Los resultados de esta prueba son:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	86.20%	95.00%	96.50%	97.50%	98.00%	98.00%
Música instrumental	38.00%	48.10%	51.30%	53.30%	54.60%	58.20%
Música con voz	78.30%	68.10%	72.00%	68.70%	69.00%	69.60%
Total	66.02%	69.51%	72.38%	72.46%	73.20%	74.67%

Tabla 5.20: *Experimento 20 (ASP 33 bases)*

5.4.6 Experimento 21

Este experimento prueba los resultados de utilizar 29 bases de proyección para extraer los parámetros ASP, obteniéndose:

Tasa de acierto (%))	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	92.00%	96.60%	96.80%	97.20%	97.70%	97.80%
Música instrumental	36.00%	48.10%	52.80%	55.50%	54.90%	57.90%
Música con voz	75.20%	70.80%	69.20%	66.90%	66.10%	66.20%
Total	66.83%	71.21%	72.41%	72.76%	72.46%	73.62%

Tabla 5.21: *Experimento 21 (ASP 29 bases)*

5.4.7 Experimento 22

Para esta prueba se utilizan parámetros ASP extraídos con 25 bases de proyección, y en estas condiciones se logran los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	86.20%	94.60%	96.10%	97.40%	97.80%	98.00%
Música instrumental	37.00%	48.70%	55.00%	53.40%	53.80%	55.00%
Música con voz	79.20%	74.80%	70.60%	71.00%	70.40%	71.30%
Total	65.90%	71.68%	73.24%	73.16%	73.24%	74.06%

Tabla 5.22: *Experimento 22 (ASP 25 bases)*

5.4.8 Experimento 23

En este experimento se utilizan 21 bases de proyección en el cálculo de los parámetros ASP, y se obtiene como resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	85.60%	93.40%	96.00%	96.60%	97.40%	97.60%
Música instrumental	35.80%	48.70%	52.50%	52.50%	52.50%	52.70%
Música con voz	79.30%	74.80%	70.60%	69.80%	69.70%	69.00%
Total	65.33%	71.27%	72.25%	72.21%	72.42%	72.38%

Tabla 5.23: *Experimento 23 (ASP 21 bases)*

5.4.9 Experimento 24

Para este experimento se utilizan parámetros ASP extraídos a partir de 17 bases de proyección, teniendo como resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	81.90%	91.80%	94.80%	95.90%	96.60%	97.40%
Música instrumental	30.60%	43.00%	48.40%	50.20%	49.50%	49.30%
Música con voz	78.80%	73.70%	69.10%	69.70%	69.00%	68.30%
Total	62.01%	68.28%	69.88%	71.07%	70.86%	70.82%

Tabla 5.24: *Experimento 24 (ASP 17 bases)*

5.4.10 Experimento 25

Se hará una última prueba, reduciendo el número de bases de proyección para extraer los parámetros ASP a 13, obteniendo estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	82.90%	92.40%	95.40%	96.50%	97.00%	97.40%
Música instrumental	32.30%	43.30%	43.80%	47.70%	46.10%	46.20%
Música con voz	77.40%	66.80%	68.90%	68.30%	69.80%	70.90%
Total	62.54%	66.52%	68.32%	69.96%	69.96%	70.45%

Tabla 5.25: *Experimento 25 (ASP 13 bases)*

5.4.11 Experimento 26

A partir de este experimento se intentarán mejorar los resultados obtenidos con parámetros ASP. Para ello se toma el número de bases que mejor resultado ha dado, teniendo en cuenta la tasa de clasificación y la reducción del número de parámetros conseguida. En este caso se toman 25 bases porque, aunque la mejor tasa se obtiene con 33 bases y 32 mezclas, con 25 bases el resultado es mejor, en general, con los modelos con menor número de mezclas. Por ello, todos los experimentos a partir de este se harán con parámetros ASP extraídos con 25 bases de proyección. En concreto para experimento se toman dichos parámetros ASP, y se le añaden los parámetros delta de éstos, obteniéndose estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	95.60%	97.80%	96.50%	97.60%	97.80%	98.00%
Música instrumental	38.00%	53.80%	58.00%	59.40%	59.40%	60.40%
Música con voz	77.00%	75.50%	74.70%	73.90%	73.30%	72.00%
Total	68.69%	74.80%	75.70%	76.31%	76.19%	76.23%

Tabla 5.26: *Experimento 26 (ASP+delta)*

5.4.12 Experimento 27

Para este experimento se añaden a los ASP, la log-energía y los parámetros delta de todas estas características. Se tienen, así, estos resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	96.80%	97.60%	96.20%	97.20%	97.80%	98.10%
Música instrumental	21.00%	26.20%	46.90%	58.50%	59.80%	60.40%
Música con voz	86.10%	86.60%	80.20%	72.70%	72.80%	72.30%
Total	65.53%	67.91%	73.20%	75.49%	76.19%	76.35%

Tabla 5.27: *Experimento 27 (ASP+logE+delta)*

5.4.13 Experimento 28

Para este experimento se toma la mejor combinación de parámetros ASP, en este caso, los ASP con sus parámetros delta (sin log-energía), que ofrece mejores resultados para

las distintas cantidades de mezclas. Una vez que se ha decidido quedarse con estos parámetros, se les aplica la normalización CMN, obteniéndose, así, los siguientes resultados:

Tasa de acierto (%)	**1 mezcla**	**2 mezclas**	**4 mezclas**	**8 mezclas**	**16 mezclas**	**32 mezclas**
Habla	95.90%	95.10%	94.20%	95.50%	97.10%	97.50%
Música instrumental	32.10%	53.70%	47.70%	56.80%	57.80%	60.90%
Música con voz	70.60%	49.50%	76.30%	68.90%	76.00%	73.20%
Total	64.47%	66.02%	71.64%	73.16%	76.23%	76.64%

Tabla 5.28: *Experimento 28 (ASP+delta+CMN)*

5.4.14. Comparativa de los experimentos con ASP

La siguiente figura muestra una comparación entre los resultados obtenidos con parámetros ASP extraídos a partir de diferentes cantidades de bases de proyección.

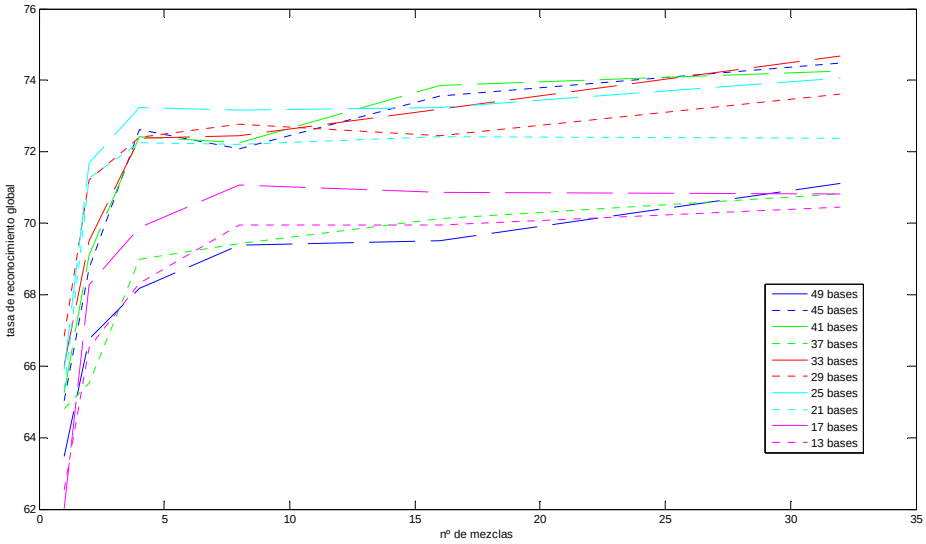

Figura 5.4: *Evolución de la tasa de reconocimiento global frente al número de mezclas, en los experimentos del 16 al 25*

De la comparativa de la figura 5.4 se puede concluir que la mejor elección puede ser utilizar 25 bases, ya que es el mínimo número de parámetros que proporciona buenos resultados, de hecho hasta con 8 mezclas es la cantidad que mejor responde, y para mayores números de mezclas no está muy por debajo de la mejor opción nunca, ni siquiera difiere en más de un punto porcentual.

Finalmente, la siguiente figura muestra los resultados de los experimentos 26 (ASP con deltas), 27 (ASP con log-energía y deltas) y 28 (ASP con log-energía, deltas y CMN). En todos los casos, se utilizan parámetros ASP extraídos a partir de 25 bases de proyección.

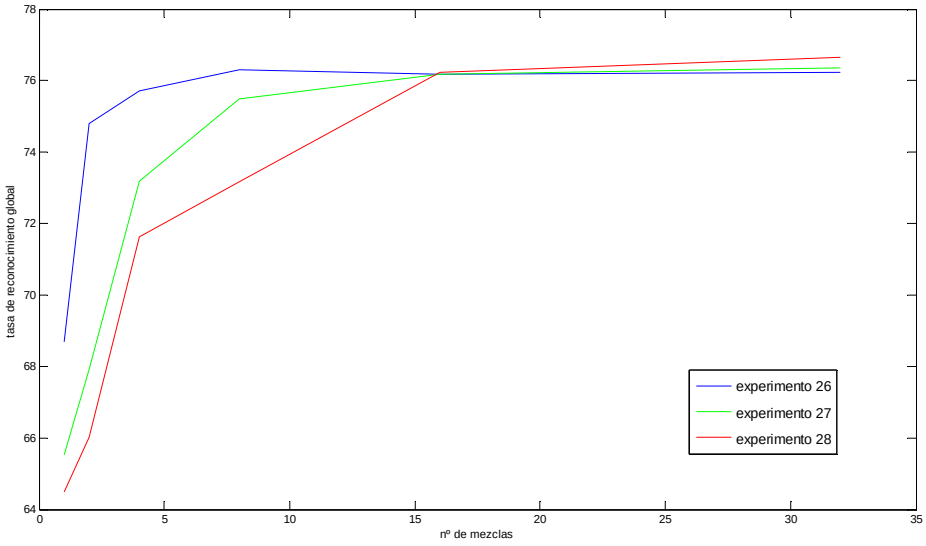

Figura 5.5: *Evolución de la tasa de reconocimiento global frente al número de mezclas, en los experimentos 26, 27 y 28*

Como se ve en la figura 5.5, a partir de 16 mezclas los resultados obtenidos son muy similares, casi idénticos. Sin embargo, con menos de 16 mezclas destacan los resultados del experimento 26 (ASP con deltas), que son siempre superiores a los de los demás. Incluso, con 8 mezclas, es superior a los otros experimentos con mayor número de mezclas.

Al igual que en el caso de los parámetros ASP, la normalización no introduce mejoras en la tasa de reconocimiento del sistema global.

5.5 Comparación entre los resultados de parámetros MFCC, ASE y ASP

En este apartado se comparan los resultados obtenidos con unas y otras características, y de esa comparación se extraerán las conclusiones que lleven al fin del proyecto, que no es otro que determinar, cuales de todas estas características de la señal de audio permiten una mejor discriminación en 3 clases, como son: habla, música instrumental, y música con voz.

Se da la circunstancia de que para todo tipo de parametrización, la clase que mejores resultados de clasificación logra, es siempre el habla. De las otras dos, la música con voz da mejores resultados que la instrumental con los parámetros ASP y MFCC, mientras que con los ASE ofrecen resultados similares. Sin embargo, para MFCC, la cosa cambia si se aplica la normalización CMN, que hace que mejoren de forma importante los resultados de clasificación de música instrumental situándola en valores claramente mayores aún que los de la música con voz. En estas condiciones cuando se logran las mejores tasas de reconocimiento globales, para toda clase de audio.

En la figura 5.6 se puede ver una comparación entre los mejores experimentos de cada tipo de parámetros.

A la vista de la gráfica queda muy claro cuáles de estos tipos de parámetros son los mejores. Los MFCC destacan sobre los demás con bastante claridad, obteniendo muy buenos resultados. Funcionan bien con las tres clases de audio a clasificar, en ninguna de ellas la tasa de acierto baja del 70%. En el experimento 8, que es el que mejor funciona, se utilizan 26 parámetros (los 12 MFCC, 1 de la log-energía, y los 13 parámetros delta de los anteriores).

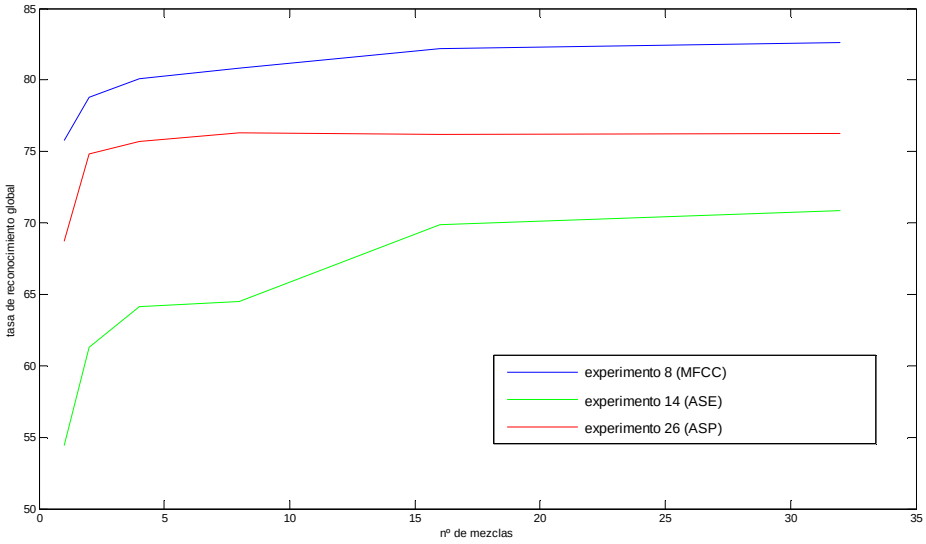

Figura 5.6: *Evolución de la tasa de reconocimiento global frente al número de mezclas, en los experimentos 8, 14 y 26*

Los parámetros ASP funcionan peor, y en gran medida se debe a que el porcentaje de acierto cuando se trata de música instrumental baja hasta un 60% en el mejor de los casos, que es sensiblemente menor al 70% obtenido con MFCC. Las otras dos clases de audio mantienen alrededor de los valores obtenidos con MFCC. En el experimento 26, el mejor con parámetros ASP se utilizan 48 coeficientes (los 24 de las bases excluyendo el que se correspondería a la log-energía, y otros 24 de sus parámetros delta).

Y los parámetros, que sin duda funcionan peor de todos son los ASE, que hacen que los resultados sean peores en todas las clases de audio a clasificar. En el experimento 14, el mejor con parámetros ASE, se emplean 100 coeficientes (49 de los ASE, la log-energía, y los 50 parámetros delta de los anteriores).

Así que no queda duda de cuáles son los mejores parámetros, porque se produce la circunstancia de que los que mejores resultados obtienen son los que menos coeficientes necesitan (y por tanto tienen menor carga computacional), y viceversa. Por lo que la mejor parametrización es la que se hizo en el experimento 8, utilizando MFCC, log-energía, y sus parámetros delta.

6 Conclusiones y líneas futuras

En este capítulo se repasa todo lo que se ha observado durante el desarrollo del proyecto, y sobre todo se extraen conclusiones a partir de los resultados experimentales que se han ido mostrando a lo largo del capítulo anterior. También se mostrarán algunas líneas de investigación que podrían servir para trabajar en ellas en un futuro.

6.1 Conclusiones

Como se dice al final del capítulo anterior, la mejor parametrización, por los resultados obtenidos de los experimentos, es la que utiliza parámetros MFCC, log-energía, parámetros delta, y todo ello normalizado con CMN. Esto quiere decir que las técnicas que llevan ya tiempo utilizándose para parametrizar voz, especialmente en aplicaciones de reconocimiento de hablante y de habla, se adaptan bien a los demás tipos de audio que nos hemos propuesto caracterizar.

La base de las parametrizaciones que se han probado han sido los parámetros MFCC, los ASE y los ASP. Estos dos últimos son parte del estándar MPEG-7. A estos parámetros base se les ha aplicado otra serie de características que se intuía que podían mejorar los resultados obtenidos con estos básicos. Esas características han sido la log-energía (energía a lo largo de la señal en escala logarítmica) y los parámetros delta (primera derivada de los restantes). También se ha aplicado en algunos experimentos la normalización CMN (*Cepstral Mean Normalization*).

La evaluación de los resultados que estos rasgos acústicos producen, nos indica que los que mejor caracterizan la señal de audio, para su posterior clasificación en las clases de

audio propuestas (habla, música instrumental, y música con voz), son los parámetros MFCC.

También se revelan resultados positivos para el uso de la log-energía, y de los parámetros delta, que mejoran los resultados. La normalización CMN no resulta muy bien con los parámetros de MPEG-7, pero, en cambio, en combinación con los MFCC, produce una mejoría muy sensible en los resultados, especialmente en la clase que peores tasas de clasificación presentaba, que era la de música instrumental.

Queda mucho por mejorar en cuanto a la caracterización de la señal musical, ya que los resultados para la clasificación del habla son ampliamente superiores a los que se producen con la música de ambos tipos. Las parametrizaciones del estándar MPEG-7 se revelan especialmente negativas para música instrumental, que apenas superan en el mejor de los casos el 60% de efectividad en la clasificación. Con MFCC tampoco son mejores en principio, pero el aplicar la normalización CMN, hace que se pase de tasas de reconocimiento inferiores al 60%, a unas algo superiores al 80%.

Para la música con voz el peor tipo de parámetros son los ASE, con los que obtiene tasas de reconocimiento que se mueven entre el 50% y el 60%. Mientras tanto, con ASP y MFCC, la música con voz llega a tasas de reconocimiento, en los mejores casos, de alrededor del 70%.

La clase de audio que siempre obtiene buenos parámetros, superiores a las otras siempre, y que rondan el 90% en muchos casos y casi el 100% en algunos, es el habla, que funciona de forma similar con las distintas parametrizaciones utilizadas, y es que claro está, resulta más fácil distinguir entre una señal de habla y cualquiera musical que entre las propias musicales. Así pues sería necesario desarrollar parametrizaciones más adecuadas para la música que es la que presenta un funcionamiento claramente inferior.

Dentro de las parametrizaciones con coeficientes ASE, cabe destacar que la resolución de octava que mejor resultado da es la de 1/8. Y para los parámetros ASP, los que mejor resultado dan, de los que se han probado, son los de 25 bases de proyección. Resulta también llamativo que para los ASP, la introducción de nuevas características (log-energía y parámetros delta) no supone un cambio significativo. En los ASE sí que

producen una mejoría. La normalización CMN produce un efecto negativo en las parametrizaciones ASE y ASP, mientras que, como ya se ha dicho, mejoran los resultados para los MFCC.

Resulta llamativo que las parametrizaciones que mejores resultados ofrecen (las que utilizan MFCC) son las que menos coeficientes necesitan, y que son las que requieren una menor carga computacional y de memoria para su cálculo. El experimento 8 (MFCC con log-energía, deltas y CMN), que es el que mejor funciona, requiere 26 coeficientes, por los 48 del experimento 26 (ASP con 25 bases y deltas), que es el que mejor funciona con ASP, y los 100 del experimento 14 (ASE con log-energía y deltas), que es el que mejor funciona con ASE. Este dato deja más claro aún que la parametrización que se buscaba en este proyecto es la del experimento 8, ya que además de dar los mejores resultados, necesita menos coeficientes de parametrización.

6.2 Líneas futuras de investigación

La clasificación de audio es un campo de investigación en el que aún queda mucho por hacer. Es una aplicación de creciente importancia debido a la gran cantidad de información, y por supuesto también de archivos de audio que se mueven a través de internet. También los equipos domésticos de reproducción de audio cada vez almacenan más archivos, y podría ser útil diferenciarlos automáticamente.

Se señalan a continuación posibles investigaciones que se podrían llevar a cabo, que guardan cierta relación con este proyecto:

- Trabajar en una mejor parametrización de la señal musical que se muestra mucho más complicada que la de voz, y en este proyecto se han logrado resultados claramente mejorables. En este sentido, se podría experimentar con otros descriptores de bajo nivel de MPEG-7 como son los relacionados con la frecuencia fundamental y la armonicidad.

- Investigar tareas más complicadas como puede ser el reconocimiento de los estilos musicales, para lo que se necesitaría una biblioteca de clases más amplia, y una parametrización más compleja.

- Buscar una normalización que mejore los resultados con los parámetros ASE y ASP, ya que la CMN ha demostrado no ser buena con estos coeficientes.

- Introducir archivos ruidosos en la base de datos y trabajar para una buena discriminación entre las clases en ambiente ruidoso.

Anexos

Anexo 1

Presupuesto del proyecto

En este apéndice se justifican los costes totales de la realización de este proyecto. Estos costes vienen dados por gastos de personal y material, y están reflejados en las tablas A.1, y A.2.

La tabla A.1 muestra las fases del proyecto con la duración de cada una de ellas. Así tenemos que el tiempo dedicado por el proyectando ha sido de 780 horas, de las que se estima que el 15% han sido compartidas con la tutora del proyecto, lo que hace un total de 897 horas. Si según la tabla de honorarios del Colegio Oficial de Ingenieros Técnicos las tarifas son de 60 €/hora, el coste de personal es de 53820 €.

La tabla A.2 muestra los costes de material desglosados en equipo informático, local de trabajo, documentación y gastos varios no atribuibles (material fungible, llamadas telefónicas, conexión a internet, transporte…). Estos gastos ascienden a 2320 €.

El presupuesto total se muestra en la tabla A.3.

Fase 1	Documentación	250 horas
Fase 2	Desarrollo de programas	100 horas
Fase 3	Experimentación y prueba	280 horas
Fase 4	Redacción de la memoria del proyecto	267 horas

Tabla A.1: *Fases del proyecto*

Ordenador	1000 €
Local (6 meses a 120 €/mes)	720 €
Documentación	200 €
Gastos varios	400 €

Tabla A.2: *Costes de material*

Concepto	**Importe**
Costes personal	53820 €
Costes material	2320 €
Base imponible	56140 €
I.V.A. (16%)	8984.4 €
TOTAL	65122.4 €

Tabla A.3 : *Presupuesto*

Bibliografía

[1] "MPEG-7 Sound-Recognition Tools". Michael Casey. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, Nº 6, Junio 2001

[2] "Analysis of the Data Quality of Audio Descriptions of Environmental Sounds". Dalibor Mitrovic, Matthias Zeppelzauer, Horst Eidenberger. Vienna University of Technology. Institute of Software Technology and Interactive Systems

[3] "Implementación de un verificador de hablante independiente del texto para dispositivos móviles Symbian", Felipe Díaz Frutos, Proyecto Fin de Carrera, Universidad Carlos III de Madrid, 2007.

[4] "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". Douglas A. Reynolds, Richard C. Rose. IEEE Transactions on Speech an Audio Processing, Vol. 3, Nº 1, Enero 1995

[5] "Study of MPEG-7 Sound Classification and Retrieval". Hyoung-Gook Kim, Edgar Berdahl, Thomas Sikora. Communication Systems Group, Technical University of Berlin, Germany

[6] "Detection of goal event in soccer videos". Hyoung-Gook Kim, Steffen Roeber, Amjad Samour, Thomas Sikora. Department of Communication Systems, Technical University of Berlin

[7] “The HTK Book (for HTK Version 3.2.1)”. Steve Young, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland. Diciembre 2002

[8] “Redes Neuronales”. Ana Bollella

[9] “A tutorial on Support Vector Machines for Pattern Recognition”. Christopher J. C. Burges. Kluwer Academic Publishers, Boston

[10] “Clustering with Gaussian Mixtures”. Andrew W. Moore. School of Computer Science. Carnegie Mellon University

[11] “MPEG-7 Overview (version 10)”. José M. Martínez. ISO/IEC JTC1/SC29/WG11 N6828. Palma de Mallorca, Octubre 2004

[12] “Evaluation and Modification of Cepstral Moment Normalization for Speech Recognition in Additibe Babble Ensemble”. Roberto Togneri, Aik Ming Toh, Sven Nordholm.

[13] “Support Vector Machine”. Wikipedia, the free encyclopedia